Analysis of the microbiome and host-transcriptome in psoriasis and atopic dermatitis

Muirhead, Gareth Sion

Awarding institution:
King's College London

ANALYSIS OF THE MICROBIOME AND HOST-TRANSCRIPTOME IN PSORIASIS AND ATOPIC DERMATITIS

A thesis submitted to King's College London
for the degree of Doctor of Philosophy
in the Faculty of Natural and Mathematical Sciences

August 29, 2017

Gareth Muirhead
Department of Informatics

Acknowledgements

First and foremost, I would like to thank my primary supervisor Sophia Tsoka. Her relentless support throughout these four years has been incredible, as was her willingness to contribute her time and insights into the many challenges I faced along the way. Her ability to stimulate scientific discussion has been inspirational to me.

I would also like to thank members past and present of the Tsoka group who have offered their insights and unreserved support during the good times as well as the hard, including Sam Neaves, Jonathan Cardoso, Laura Bennett and Aristotelis Kittas.

Special praise is required for my second supervisor Frank Nestle. His guidance and immense knowledge of skin disease bolstered my biological background and stimulated many avenues of research. I would further like to extend a big thank you to all members of the Nestle group for the many thought-provoking discussions which have really contributed towards my understanding of the underlying biology.

Without the MAARS consortium, this thesis would have not been possible and I express my utmost gratitude to the whole group. The inspiring discussions and meetings we had over the years were truly fruitful and I am incredibly grateful to have been part of the team. There are many people who have contributed including the clinicians and laboratory technicians who collected and generated the datasets, as well as the rich pool of immunological expertise including Harri Alenius, Bernhard Homey, Nanna Fyhrquist and Frank Nestle to name just a few.

The MAARS bioinformatics and analysis group has been instrumental during my time here including Vassili Soumellis, Björn Anderrson and Sophia Tsoka, with special thanks to Marine Jeanmougin, Stefanie Prast-Neilsen, Max Bylesjo and Mauricio Barrientos-Somarribas.

Their collaborations and in-depth discussions regarding the analysis of this challenging dataset have been exceptional.

I also extend a sincere thanks to the many collaborators who have helped me along the way including Davide Pennino, Emanuele de Rinaldis, Helen Alexander, Lazaros Papageorgiou and Lingjian Yang. To my friends at King's College, with special mention to Lukas Diekmann, Christopher Hampson and Martin Chapman who have been with me for the ride, you have been the source of many unforgettable memories. Emily McLean has not escaped my notice and has provided nothing but unyielding support.

Finally, I would like to thank my parents Carys and James, as well as my brother Chris for their words of encouragement and support that have never wavered.

Abstract

Skin is the primary interface to the external environment, protecting us from harmful compounds and infection. Over many millennia, coevolution has culminated in a mutualistic relationship between ourselves and a specialised, yet diverse resident community of microorganisms. This microbiota is thought to exist in a delicate balance to maintain homeostatic equilibrium, and plays a role in the shaping of our immune system. Skin diseases are some of the most common human disorders and present a considerable economic burden. There is now growing appreciation of the role the cutaneous microbiome plays in disease, and how host-microbe interplay is associated with disorders of the immune system. The cutaneous microbiota as well as the host transcriptome and the interactions between them is the focus of this thesis.

Computational methods were used to examine the microbiota and transcriptomes of a large cohort of matched samples from healthy volunteers and patients with Atopic Dermatitis (AD) and Psoriasis (PSO). The community composition of inflamed and healthy skin was assessed and the specific species which are over-represented on diseased skin were identified. In parallel, the host gene expression was profiled to identify common and disease specific transcriptional signatures revealing clinically relevant pathways.

Next, using schemes of dimensionality reduction and computational methods, the associations between microbes, disease severity and host transcription was interrogated. *Staphylococcus aureus* demonstrated an impressive relationship with AD associated gene signatures, whereas host-microbe associations in PSO were inconclusive. Using Weighted Gene Co-expression Networks Analysis (WGCNA), gene-gene interaction networks were reconstructed and differentially connected modules between healthy and inflammatory states were identified. Modules encoding processes for the epidermal barrier, extracellular matrix, non-coding RNA metabolism and immune system processes were all associated with

the relative abundance of *S. aureus*.

Overall, whilst associations in psoriasis were inconclusive, a range of host-microbe interactions were uncovered in AD. The results presented in this thesis contribute towards a greater understanding of the differences in the cutaneous microbiome and the transcriptional mechanisms which underlie allergic and autoimmune inflammation.

Abbreviations

AD ..... Atopic Dermatitis

ADL ..... Atopic Dermatitis Lesional

ADNL ..... Atopic Dermatitis Non-Lesional

AMP ..... Antimicrobial Peptide

AUC ..... Area Under the Curve

BC ..... Bray Curtis

BICOR ..... Biweight Mid-correlation

CTRL ..... Control

DC ..... Dendritic Cell

DEG ..... Differentially Expressed Gene

ECM ..... Extracellular Matrix

EDC ..... Epidermal Differentiation Complex

FDR ..... False Discovery Rate

FWER ..... Family-wise Error Rate

FLG ..... Filaggrin

GSEA ..... Gene Set Enrichment Analysis

IgE ..... Immunoglobulin E

IL ..... Interleukin

INF ..... Interferon

IPA ..... Ingenuity Pathway Analysis

LCE ..... Late Cornified Envelope

LFC ..... Log_2 Fold Change

NMDS ..... Non-metric Multidimensional Scaling

MAARS ..... Microbes in Allergy and Autoimmunity Related to the Skin

MHC ..... Major Histocompatibility Complex

ORA ..... Over-representation Analysis

OTU ..... Operational Taxonomic Unit

PCA ..... Principal Components Analysis

PCoA ..... Principal Co-ordinates Analysis

PSO ..... Psoriasis

PSOL ..... Psoriasis Lesional

PSONL ..... Psoriasis Non-Lesional

QIIME ..... Quantitative Insights Into Microbial Ecology

rRNA ..... Ribosomal Ribonucleic Acid

TEWL ..... Trans-epidermal Water Loss

Th ..... T-helper

TNF ..... Tumor Necrosis Factor

WGCNA ..... Weighted Gene co-expression Network Analysis

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Our interactions with the world are experienced through the vast $2m^2$ of cutaneous membrane that encapsulates our body. Whilst this primary interface protects us from the harmful agents present in the external environment, it is also the home of a specialised, yet diverse resident community of microorganisms [1]. This microbiota has accompanied us throughout our evolution culminating in a mutualic kinship where they are supported by the nutrient rich supply of secreted sebum and shed squame and in return we obtain protection from potentially offending microbiota. Sometimes described as the forgotten organ [2], the totality of the cells that make up the microbiota equals that of the human [3]; however, unlike many other organs, the microbiota is dynamic and demonstrates a considerable amount of diversity between individuals [1]. Enteric dysbiosis, when the community composition deviates from a healthy ecological equilibrium, has now been linked to inflammatory disease [4, 5, 6], although much less is known about the relationship between between skin microbiota and inflammatory skin pathologies. The purpose of this thesis is to explore this relationship between cutaneous dysbiosis and host-transcriptional disease signatures.

1.1 Studies of the microbiome

In the fifteen years since the initial draft of the human genome [7], incredible advances have been made in both sequencing technologies and the bioinformatics capabilities to analyse high-throughput data. Research showed that a 1500bp sequence encoding for a subunit of the prokaryotic ribosome, known as the 16S rRNA gene, has a slow rate of evolution and acts as molecular chronometer [8]. It was found that mutations within this gene can

be used as a diagnostic fingerprint to distinguish between taxonomy up to the species level.

The utilisation of sequencing technologies for the 16S rRNA gene has made it possible to explore the human microbiota at resolutions previously infeasible by phenotypic methods [8]. Ever since, 16S sequencing has been used to explore the microbiota and has revealed many novel commmensal species occupying the human body which were previously unknown and unculturable [9]. We are now in the midst of a metagenomic revolution and beginning to establish the fundamental principles that govern the relationship between the human organism and its microbial cohabitants.

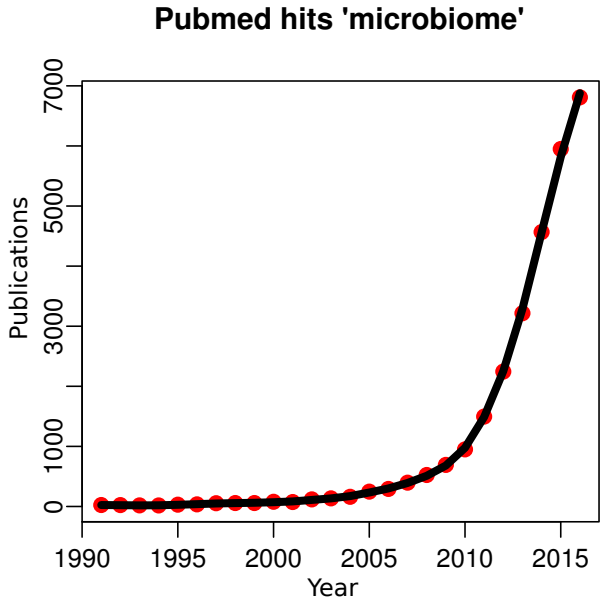

Figure 1.1: Pubmed hits for the query 'microbiome'

Interest in the microbiome and its role in health is rapidly increasing. A search for the term 'microbiome' in Pubmed clearly indicates the excitement in this new area of research where the numbers of associated publications have rapidly increased in recent years (**Figure 1.1**). A popular database for storing metagenomic data, MG-RAST [10], currently stores over 250,000 metagenome samples. Studies which have profiled the 16S rRNA gene are ranging and varied. There are now microbiome surveys for several human habitats including the

eyes [11], the blood [12], and the lungs [13].

1.2 The gut microbiome

The overwhelming majority of research into the human microbiome has been performed in the gut which is mostly dominated by the *Firmicutes* and *Bacteroidetes* phyla [14, 1, 15]. As demonstrated by a recent metagenomics analysis, the genetic capability of gut microbiome is vast and consists of at least 150-fold more microbial genes than the human genome [15].

The community of microbiota is believed to be acquired at birth; however, distinct differences in composition can be observed between babies who were delivered by cesarean or born vaginally [16]. Genetic studies have indicated that components of the microbiome are heritable suggesting that host genetics help to shape the microbiome. It has been shown that the biotic composition of twins is more similar than unrelated individuals, and the composition between monozygotic twins is more similar than dizygotic twins [17]. Many environmental factors affect the gut microbiome; one of which includes the host diet [18]. Longitudinal studies of the infantile microbiota have shown that changes to diet such as the weaning from breast milk to solid foods can be detected by shifts in community composition [19]. Changes in the gut microbiome have also been observed with respect to body mass index (BMI) as well as gender [20] and, as would be expected, antibiotic treatment [19].

It is now becoming increasingly clear that the microbiome plays a critical role in the shaping of our immune system [21, 22]. Studies in mice which have never encountered microbiota, known as germ-free mice, are characterised by reduced numbers of immune cells which are functionally compromised and the mice themselves are more susceptible to infections [23]. One theory suggests that under-stimulation of the immune system during developmental periods results in susceptibility to immune dysbalance [24]. This view is known as the 'hygiene hypothesis' [25] and reflects the increased prevalence of inflammatory diseases amongst industrialised countries. This is further supported in that babies delivered by cesarean, as well as those treated with antibiotics experience an unrecoverable reduction in microbiota diversity [26]. Improved standards of hygiene due to clean water, sanitation as well as continuous hand sterilisation have seen rates of infection plummet; however, this

rapid change in lifestyle may result in an ill-prepared and dysfunctional immune system [27].

The overwhelming consensus view is that the gut microbiome plays an essential role in host health. The genetic capability provided by the microbiome aids in metabolism of indigestible compounds, training of the host immune system, and providing resistance to infection [22]. If the microbiome behaves as a functional organ, then it is also likely to be susceptible to disease. The intimate relationship with our microbiota exists in a delicate homeostatic equilibrium and when this is disturbed, a state of dysbiosis can occur. Shifts in community composition have been linked to wide range of pathologies including Crohn's disease [4, 28, 6], obesity and inflammatory bowel disease [5, 29]. As dysbiosis is clearly a factor, there is now increasing interest regarding the therapeutic potential for modulation of the microbiome.

1.3 The skin

The skin accounts for approximately 16% of total body weight and is our first line of defence against foreign bodies. The healthy cutaneous membrane plays critical roles in the maintenance of temperature, lipid and vitamin D synthesis as well as regulation of appropriate fluid loss and retention. The skin is mostly composed of two structures, the epidermis, which is the main barrier, and the dermis which supports the epidermis by providing structure and facilitating nutrient transport. [30]

1.3.1 Structure of healthy skin

The dermal compartment is beneath the epidermis and contains a vast extracellular matrix (ECM) as well as the hair follicles and subaceous glands. Two main layers compose the dermis; the papillary layer, and the reticular layer. The reticular layer consists of a vast amount of irregularly connected collagen and is responsible for structural integrity. The papillary layer is the upper most layer of the dermis. This layer contains sensory neurons, capillaries, lymphatic vessels as well as epidermal ridges which protrude into the epidermis and facilitate the transfer of products, between the epidermis and dermis as well as structural support. [30]

The epidermis consists of several keratinocyte layers. The layer superficial to the dermis

is the stratum germinativium which consists of a high abundance of basal cells which differentiate to replace dead cells which are shed from the outermost layers. When basal cells divide, the daughter cell moves up a layer into the stratum spinosum which consists of keratonocytes bound together by desmosomes, a structure which anchors cells together. The stratum spinosum also contains a rich source of Langerhans cells, which are specialised dendritic cells situated in the cutaneous membrane and mediate immune responses against offending microbiota. As keratinocytes continue to divide from the basal layers, they reach the stratum granulosum, where they begin to produce high quantities of keratin protein. The final layer, known as the stratum corneum consists of many keratinocyte layers, which have been hardened and become flatter due to the volume of keratin production. Keratinisation, also known as cornification, occurs in this layer and is the formation of a strong protective barrier of dead keratinocytes bound together by special desmosomes known as corneodesmosomes. Keratinocyte layers in the the stratum corneum remain in this layer for approximately 1-2 weeks when they are shed. They are then continuously replenished by younger cells pushed up from the subficial layers. [30]

1.3.2 The skin microbiome

The skin is home to a diverse myriad of resident microbiota including fungi, viruses, mites and bacteria [31]. It is hypothesised that this resident microbiota may provide resistance to infection by production of antimicrobial peptides in exchange for host supplied sebum [32]. This diverse community is believed to be acquired at birth and closely resembles the mother's vaginal microbiota [16]. The landscape of microenvironments in the skin is diverse and is defined by differences in pH, sebum, moisture levels and temperature, requiring the microbiota to be resilient to environmental changes [33].

One of the major functions of the skin is to regulate temperature. Sweat glands are unevenly distributed across the body resulting in diverse skin microenvironments. Several different types of sweat gland are present on the skin including eccrine, apocrine and apoeccrine as well as sebaceous glands which secrete sebum; all of these contribute to mixtures of varying compositions supplying nutrients for microbial metabolism [31]. Sweating is controlled by the nervous system and responds to changes in temperature, physical activity, emotions and stress [34]. Differences in moisture and nutrient composition due to sweating as well as anatomic variation such as hair follicle density all contribute to skin microenvironments which may be colonised by niche-specific microbiota [34, 31].

Eccrine sweat is 99% water and includes amino acids as well as the antimicrobial peptide dermacidin [35]. They are found across the entire body with high density on the palms and soles with only few on the back [34]. Apocrine sweat glands are found in the perianal regions, armpit and eyelids. The secretion from apocrine glands is an odourless substance rich with lipids and proteins, however, this is metabolised by bacteria such as *Corynebacteria* into volatile odorous compounds and is associated with bromhidrosis [31, 34]. Little is known about apoeccrine glands but their secretion is thought to be similar to eccrine glands and they occur mostly in haired areas [34]. Sebaceous glands secrete an oily compound called sebum consisting of fatty acids, wax esters and cholesterol. They are found in high density on the face and are closely related to the skin condition acne. The sebaceous gland is directly attached to the hair follicle in a structure known as the pilosebaceous unit which are anoxic and are colonised by microbiota including *Propionibacterium acnes* [36, 31].

The distribution of eccrine, apocrine and sebaceous glands is variable across the human body and several studies have attempted to map the cutaneous microbiome by dividing up skin locations into moist, dry and oily sites. Moist sites contain an abundance of eccrine and apocrine glands and consist of flexural areas such as behind the knee, elbows and between the toes [37]. These regions have been found to be colonised mostly by *Corynebacterium* (*Actinobacteria*) and *Staphylococcus* (*Firmicutes*) [38]. Analyses of body sites have shown that the diversity of the skin microbiota varies according to the body site sampled [39]. The oily sites have been found to have the lowest diversity amongst skin sites in terms of both metabolic and taxonomic diversity [33, 38, 40]; however, they are enriched for specific metabolic pathways such as glycolysis [33] indicating that taxa adapt to their skin-specific niche. The taxa most dominant within the oily regions include *Propionibacteria* (*Actinobacteria*) which colonise sebaceous glands and *Staphylococci* (*Firmicutes*) [38]. The dry areas consist of fewer sebacous glands and are dominated by *Betaproteobacteria* [38]. Taxa residing in dry sites also showed niche specific genetic capability through over-representation of genes involved in citrate cycle [33]. Overall, these studies have shown that community composition is variable between body sites and further indicates that specific species are adapted to their microenvironments via the over-representation of specific metabolic pathways.

Whilst body site is seen to be the major factor which shapes the microbiome, several other factors are also associated with community composition, one of which is the human environment. A study by Ying et al. [40] showed compositional differences between people residing within urban and rural areas of the same city. Furthermore, the same study showed a relationship with age in which adults have more diverse microbiomes than both adolescents and elderly, and elderly people were found to have lower abundance of *Propionibacterium* [40]. Little is known about the effect of clothing, soaps and other cosmetics on the skin microbiota although it plausible that factors of this kind may influence community composition [31].

When comparing the skin microbiota to different human body sites, the skin is compositionally dissimilar to that of the gut [39, 1] and has a higher viral and fungal component than other body sites [33]. The human microbiome project found that the skin had comparable within-site diversity (α-diversity, see Section 2.1.2.2) to other (non-skin) body sites, however, analysis of between-site diversity (β-diversity) showed that some skin sites are amongst the most diverse microbiomes in the human body [1]. This finding has also been supported by others [39] and indicates that the skin microbiota may be less stable between individuals.

Costello et al. [39] showed that microbiota composition was more stable within an individual over time than between individuals. Furthermore, a recent study of the hand microbiota showed that only 13% of species are shared between any two individuals [41]. Given the high degree of inter-individual variability, the idea of a 'personalised microbiome' has been proposed. Subsequent works have built upon this idea and demonstrated that residual skin microbiota left on objects such as keyboards can be used as a personal signature for forensic analysis [42].

There is now an increasing drive to determine the relationships between host and microbiome in the context of health and disease. It is well known that dysbiosis or loss in species diversity is a common symptom of gut inflammatory diseases, and it has been postulated that this may exacerbate inflammation [22]. The cutaneous membrane is a rich source of immune cells including T cells, dendritic and mast cells and recent studies in mice have shown that the skin microbiota is associated with IL-1 signalling demonstrating a host immune system-microbiome interplay [43]. Only with further analysis into host-microbiome

interactions will be able to determine the relationship between the skin microbiome and inflammatory skin diseases such as Atopic Dermatitis and Psoriasis.

1.4 Atopic Dermatitis

Atopic dermatitis is a relapsing inflammatory skin condition associated with erythematous and intensely puritic papules which become lichenified in the chronic phase [44]. The disease has a negative impact upon the quality of life and treatment for atopic dermatitis is long term and can be a financial burden. Approximately 15-30% of children have atopic dermatitis, 85% of which develop before the age of five, and 2-10% of adults [44]. The disease presents on different body sites varying with age and often affects the facial areas in infancy and mostly flexural areas in adults [45].

AD is a multifactorial disease and is believe to be associated with complex interactions between the immune system, host genetics and dysfunction of the epidermal barrier, however, its etiology is currently unknown. Genetic factors clearly play a role in AD as the concordance rate amongst monozygotic twins is greater than dizygotic twins [46]. Furthermore, the risk of developing the disease is twice as high if at least one parent has AD [44].

The condition is associated with increased immunoglobulin E (IgE) sensitisation to environmental allergens such as dust mites and is considered to be the first component of the 'atopic march' or the 'atopic triad' in which half of sufferers develop asthma, and two thirds will acquire rhinitis [47]. It is important to note that whilst the majority of patients have elevated IgE serum levels, some patients do not and this has resulted in definition of two AD subtypes. The major most prevalent subtype is known as extrinsic AD and accounts for 70-80% of patients and is associated with IgE sensitisation. The second subtype have normal serum IgE levels and is known as intrinsic AD or non-atopic which have reduced IL-4 and IL-13 levels and accounts for 15-30% of patients [48].

The most significant findings from genetic studies have linked dysfunction of the epidermal barrier to the pathogenesis of AD. Studies have shown that many patients with AD carry loss of function mutations within the filaggrin (FLG) gene which is a key component of the statum corneum [49, 50]. FLG aggregates keratin filaments in the epidermis and plays a role in modulation of epidermal pH as well as keeping the skin barrier hydrated

through generation of natural moisturising factors [51, 50]. Thus, FLG loss of function mutations are thought to be related to the dryness often characteristic of AD lesions [49]. With a defective barrier, it is thought that increased barrier permeability results in elevated transepidermal allergen transfer which increases the contact between environmental allergens and immune cells [52].

AD is considered to be a biphasic disease which is dominated by T helper (Th) 2 cells in the acute phase, and Th1 cells in the chronic phase [44]. In the acute phase, environmental allergens enter the epidermal barrier and are processed by skin-resident dendritic cells (DCs). Upon an encounter, DCs mature and down regulate E-cadherin allowing detachment from neighbouring keratinocytes in conjunction with up-regulation of major histocompatibility complex (MHC) class I & II molecules [53]. Keratinocytes in AD lesions express high levels of thymic stromal lymphopoietin (TSLP) which condition DCs to induce Th2 cell differentiation [54, 53]. The detached and activated DCs migrate to the lymph node and present the antigen on MHCII which is recognised by T cell receptors (TCR) on naive CD4+ T cells promoting the polarisation of Th2 cells. [53]

Upon return to the skin, Th2 cells secrete IL-4, IL-5 and IL-13 [53, 55]. IL-5 promotes proliferation and activation of eosinophils, whereas IL-4 and IL-13 induce class switching in B cells stimulating the production of antigen-specific IgE [56]. This primes mast cells which, upon cross-linking with an allergen, degranulate releasing proinflammatory molecules. Besides inducing B cell class-switching, Th2 cytokines inhibit the expression of terminal differentiation genes in keratinocytes such as filaggrin and loricrin which further impairs barrier integrity [56]. In the chronic phase, the inflammation shifts from a primarily Th2 mediated to a mixed response in which Th2, Th1 and Th22 cells play a role in inflammation as well as fibrotic remodelling culminating in lichenification [57].

Aside from the defects associated with genetic mutations within the FLG gene, other barrier deficiencies are present in AD resulting in xerosis and an increase of transepidermal water loss (TEWL). These include variants within the region of genes which encode the epidermal differentiation complex (EDC), as well as reduced ceramide and lipid expression [58, 51, 59, 44]. It is important to note that the uninvolved skin of patients with AD is not clinically normal and is also dry [48]. Reports have shown decreased levels of ceramides [60], as well as increased expression of Th2 products in non-lesional skin [61].

1.4.1 The AD microbiome

Early culturing studies performed in the 1970s revealed an over-representation of *Staphylococcus aureus* [62] on the lesional and non-lesional skin of AD patients. This finding has been replicated by others who showed that *S. aureus* is encountered more often, and to higher density on lesional skin than on non-lesional skin [63]. Based upon this evidence, AD is often treated with antimicrobial agents to control the growth of *S. aureus* [44], however, the contribution of the microbiome to disease is not yet fully understood. Studying the role of *S. aureus* and the further resident community can enable a better understanding of the pathogenic and protective host-microbe interactions which may trigger and exacerbate inflammation.

The most in-depth analysis of the AD microbiome utilised 16S sequencing on a cohort of 12 children with AD and 11 healthy controls [64]. Kong et al. confirmed a strong presence of *S. aureus* but also demonstrated several other disturbances to the atopic microbiome which could not have been identified using culturing methods. The study showed that the microbiota exists in a state of dysbiosis due to a loss in species diversity [64]. Furthermore, it was shown that patients with severe disease experienced further loss in diversity, and that patients undergoing treatment had significantly higher diversity than those without. A follow up study in a single patient showed that the diversity during an inflammatory flare up (flare) which was untreated was lower than that of a bleach bath treated flare. This analysis indicated that anti-inflammatory or antimicrobial therapy restores a level of diversity and may limit the effects of dysbiosis. Furthermore, *S. aureus* was found to strongly correlated with species diversity thus indicating that this pathogen is a major factor in dysbiosis [64].

The role that *S. aureus* plays in the pathogenesis of AD is an intense area of research. *S. aureus* produces superantigens including staphylococcal enterotoxin A and B [65]. It has been shown that the abundance of superantigen isolated from *S. aureus* strains on AD skin is higher than superantigen isolated from *S. aureus* strains on healthy skin [66]. Moreover, the same study also showed that superantigen was correlated with T cell activation as well as disease severity and thus may be one of the mechanisms by which *S. aureus* exacerbates or initiates inflammation.

Kong et al. [64] also reported that *S. epidermidis* was of increased abundance on atopic

skin, albeit not to the extent of *S. aureus*, and that the relative abundances of *Streptococcus*, *Corynebacterium* and *Propionibacterium* increased post-treatment restoring a level of species diversity. The role of *S. epidermidis* is an intense area of debate and is often considered to be a healthy skin commensal [31]. Studies have shown at least two possible mechanisms by which *S. epidermidis* may protect the host against over-colonisation of *S. aureus*. *S. epidermidis* is thought to selectively inhibit *S. aureus* growth by production of phenol-soluble module (PSM) peptides which have similar characteristics to host antimicrobial peptides [67]. Another study showed that *S. epidermidis* can induce host keratinocytes to produce antimicrobial peptides via the toll-like receptor (TLR) 2 signalling pathway, enabling the host to improve response against pathogens [68, 65]. Other species such as *P. acnes* which often colonises healthy skin [31], may also play a beneficial role by the metabolism of glycerol into short chain fatty acids (SCFA) including propionic acid which suppress the growth of *S. aureus* [69].

Whilst it is clear that *S. aureus* plays a role in the pathogenesis of AD, the mechanisms of action are not yet fully understood. One way to gain insight is to study host-microbe interactions and evaluate the effect of colonisation on the expression of host genes. Such analysis would allow for a greater understanding into how *S. aureus* virulence factors are associated with the expression of structural epidermal barrier genes and host immune processes [65].

1.5 Psoriasis

Psoriasis has a current prevalence of around 3% in the USA population and its incidence rate has doubled over the last 30 years [70]. The condition impairs quality of life and treatment can be a considerable economic burden. Approximately 30% of psoriasis patients also have psoriatic arthritis (PsA); these patients endure joint pains and debilitating physical impairment [71]. As well as PsA, psoriasis patients are characteristic of many comorbidities including metabolic syndrome [72], type 2 diabetes and cardiovascular disease [73]. The age of onset of psoriasis is bimodal, with the first occurring between the ages of 20-30 and the second between 50-60.

Psoriasis is a chronic, currently incurable inflammatory skin disorder associated with silvery plaques due to hyperproliferation of keratinocytes and incomplete cornification in the

statum corneum [73]. Keratinocytes in the basal layer divide at a faster rate resulting in a thickened epidermis with a defective outer layer in which the turnover time of basal cells is rapidly increased.

Involvement of the immune system in psoriasis is clear with increased infiltration of both innate cells including neutrophils and macrophages, as well components of the adaptive immune system including T cells which secrete proinflammatory cytokines resulting in inflammation [73]. Psoriatic skin contains an abundance of T-helper 1 (Th1) cytokines including IL-1, IFN-γ and TNF-α, therefore, early views considered psoriasis to be mostly associated with Th1 cells [74]. This model was challenged when it was discovered that cytokines produced by Th17 cells including IL-17A and IL-22 were prevalent in lesions as well as increased infiltrates of Th17 cells themselves [75]. IL-17A has wide spread proinflammatory effects which induces cytokine production in many other immune cells.

Whilst the trigger of psoriasis is not fully understood, it is thought that genetic factors in conjunction with environmental factors such as stress, microbiota or trauma result in stressed keratinocytes which release nucleic acids, as well as pro-inflammatory cytokines TNF, IL-6 and IL-1B [76]. The self DNA/RNA complexes with an antimicrobial peptide called LL-37 which activates plasmacytoid dendritic cells (pDCs). The activated pDCs secrete interferon-α, which along with keratinocyte derived cytokines leads to the activation of dermal dendritic cells. Once activated, dermal dendritic cells migrate to the lymph nodes where cytokines including IL-23 and IL-12 induce a Th17 or Th1 phenotype in naive CD4+ t cells by presentation of a currently unknown antigen [76, 73]. Activated Th17 and Th1 cells migrate back to the skin along a keratinocyte-derived chemokine gradient where Th1 cells secrete interferon-γ and TNF-α, and Th17 cells secrete IL-17A, IL-17F and IL-22 [76]. These cytokines drive the proliferation of keratinocytes and promote secretion of antimicrobial peptides including DEFB4 and S100A7 which are expressed to great magnitude in psoriatic lesions [77] and neutrophil chemoattractants [76, 73].

The mechanisms underlying psoriasis are not yet fully understood and the triggering event is still unclear. Genetic factors clearly play a role as genome wide scans revealed that psoriasis is associated with variants in the IL23A and IL23R genes [78]. IL-23 is secreted by dendritic cells and is involved with the differentiation of Th17 cells where variants may enhance induction of the Th17 phenotype [74]. Biologic treatments that target IL-23 have

shown effectiveness in the treatment of psoriasis [79, 80]. Another possible factor relates to impairment to the epidermal barrier. Genetic variants have been discovered in the LCE3B and LCE3C genes which may impair skin barrier function enabling increased permeability and entry of immunoreactive microbial products into the skin [81]. Several of the genetic variants associated with psoriasis are also factors in Crohn's disease [80]. Crohn's disease is thought to be associated with enteric dysbiosis indicating that microbial stimuli may well be a factor in psoriasis [80]. The basis for genetic association is further compounded by patients undergoing bone marrow transplantation (BMT) which showed resolution of psoriatic inflammation [82]. Similarly, there are also reports of patients developing psoriasis after BMT from a psoriasis affected donor indicating that psoriasis is caused by genetic factors affecting bone marrow-derived immune cells [83].

1.5.1 The PSO microbiome

Genetic mutations within the innate and adaptive immune system, as well as to genes encoding for components of the epidermal barrier suggest a potential role of the microbiome in psoriasis. Several analyses of the psoriatic microbiome have been performed, however, the results are varied between studies.

Gao et al. [84] found that *Propionibacterium acnes* was reduced on lesional skin. Further reduction within the relative abundances of *Proteobacteria* and *Actinobacteria* phyla were observed as well as an increase in *Firmicutes*. These initial findings confirmed that the composition of microbiota on lesional and non-lesional skin were perturbed in psoriasis.

A more recent study indicated that the combined relative abundances of *Corynebacterium*, *Propionibacterium*, *Staphylococcus* and *Streptococcus* were increased on lesional skin [85]. Furthermore, in the same study, psoriasis was shown to have reduced species richness on both lesional and non-lesional skin, as well as higher intra-group variability compared to control samples. In response to a high impact paper describing enterotypes in the gut microbiome [14], Alekseyneko et al. described a similar phenomenon in the skin, which they called *Cutaneotypes*. Cutaneotypes were defined as two clusters distinct microbial compositions: Cutaneotype 1 was associated with *Actinobacteria* and *Firmicutes* whereas cutaneotype 2 was associated with *Proteobacteria*. The *Firmicutes-Actinobacteria* dominant cutaneotype was significantly associated with lesional psoriasis.

A study of the psoriatic microbiota isolated from skin biopsies reached different conclusions. Fahlen et al. [86] described a trend for reduced *Actinobacteria* in psoriatic skin as well as increases in *Proteobacteria* at the phylum level. At the genus level they found reductions in *Propionibacteria* and *Stapylococcus* on lesional skin from the limb.

Overall these studies have shown that the psoriatic microbiota does vary with psoriasis status. Studies on swabs showed general trends for increases in *Firmicutes* and reductions in *Proteobacteria*, however, there were some inconsistencies. For example, Gao et al. [84] identified *Actinobacteria* as being significantly under-represented on psoriatic skin, whereas Alekseyneko et al. [85], did not report any significant difference in *Actinobacteria*, but described psoriasis samples as being associated with a '*Firmicutes-Actinobacteria*-high' cutaneotype. Analysis of biopsy data were even more inconsistent and suggests that microbiota may be able to penetrate the skin and interact with the host indicating another role for intracutaneous microbiota. Further analysis with larger sample sizes may be required to define the core characteristics of the psoriatic microbiome.

1.6 Contributions

In comparison to the gut, little is known about the skin microbiome and its role in host health and disease. Whilst there are studies evaluating the composition of the atopic dermatitis and psoriasis associated microbiomes, these studies were performed on few samples, and the results are varied, particularly in the case of psoriasis. This thesis presents an analysis of the largest inflammatory skin disease microbiome dataset to-date consisting of more than 600 samples of healthy, non-lesional and lesional skin from healthy volunteers and patients with either atopic dermatitis or psoriasis generated by the Microbes in Allergy and Autoimmunity Related to the Skin (MAARS) consortium. As well as 16S sequencing of the cutaneous microbiota, the host transcriptome was profiled from biopsies taken at the same location providing a unique opportunity to investigate host-microbe interactions. The results in this thesis contribute towards a greater understanding of bacterial dysbiosis and the relationship with host gene expression in the in context of inflammatory skin pathologies.

In **Chapter 3**, the microbiota is comprehensively evaluated using the largest skin inflammation associated 16S dataset to date. Across all levels of the phylogenetic tree, including both lesional and uninvolved cohorts, this analysis identifies the taxa which are either over-represented or under-represented in both psoriasis and atopic dermatitis. Here, whilst confirming the presence of known pathogens such as *S. aureus* in atopic dermatitis, several novel species were identified as potential pathogens in psoriasis, both on lesional and non-lesional skin. As well as differential abundance analysis, the association of species diversity with disease severity was established. Finally, using a co-occurrence network analysis of disease associated species, differences in the topology of microbe-microbe interactions were identified between AD and PSO.

In **Chapter 4**, the transcriptome is interrogated to identify differentially expressed genes which may play critical roles in the mechanisms which underlie atopic dermatitis and psoriasis. The genes which are active in the non-lesional tissue of both diseases were contrasted revealing similarities and differences in the transcriptional architecture of disease susceptible skin. A similar comparative analysis of the lesional tissue was performed and the pathways which are common to both diseases or preferentially expressed in either AD or PSO were determined. These results identify the particular cytokines and components of the immune system which were common or specific to both diseases.

In **Chapter 5**, the microbiome and host-transcriptome were integrated to identify host-microbe associations. Using differentially expressed genes and suitable schemes of dimensionality reduction, the microbe-associated host transcripts were identified. The pathways identified are highly relevant to disease pathology and provide insight into how pathogenic microbes interact with the host to drive inflammatory disease.

In **Chapter 6**, skin co-expression networks were constructed within the weighted gene co-expression network analysis (WGCNA) framework in order to identify modules of highly co-expressed genes. By performing comparative network analysis, modules which were differentially co-expressed between healthy, lesional and non-lesional tissue were identified by module preservation analysis. Two modules of genes were found to be disconnected in a healthy skin which formed a tight interacting coexpression module in a diseased state. Using the same networks, eigengenes were used to elucidate host-trait relationships. This

analysis revealed clinically relevant modules of the inflammatory transcriptome which co-varied with both disease severity and pathogen abundance providing further insight into microbe associated gene signatures.

Chapter 2

Methods to study the microbiome and transcriptome

This chapter introduces the technology and methodologies used to generate and analyse transcriptome and microbiome datasets. Following this, the methods and sampling protocols used to generate the MAARS cohort are described.

2.1 The microbiome

Early studies of microbiota were performed on a phenotypic basis [87, 9] and to identify differences between samples, morphological comparisons were performed using culturing methods. One of the major drawbacks of quantitative culturing is that many species are not easily cultured, especially anaerobic microbes with one estimate suggesting that more than 99% of species cannot be cultured by traditional means [88]. Culturing approaches are also slow, labour intensive and thus cost ineffective. To increase resolution, molecular genotyping methods were developed which are now far quicker and more accurate at identifying species than culturing approaches [89].

2.1.1 Methods to study the microbiome

2.1.1.1 16S sequencing

With advances in DNA sequencing technologies, modern approaches emerged for bacterial classification based upon comparative genomics. The first molecular markers focused around the ubiquitous 16S rRNA gene [8] which is a highly conserved critical component of

the translational machinery. The 16S rRNA gene is a short 1500bp sequence and behaves as a molecular chronometer [90]. As the molecular function of the 16S gene is necessary for cellular operations, it is one of the most conserved genes with a very low rate of mutation [89].

The 16S rRNA molecule has several characteristics which make for an ideal phylogenetic marker. First, the sequence is present in all prokaryotic organisms and contains both variable and highly conserved regions [90]. The conserved regions mutate slowly and are targeted by universal PCR primers enabling global amplification of all 16S genes contributed by the bacterial population. Nine 'hyper-variable' regions evolve faster, and it is within these regions that sequence heterogeneity is observed between bacteria reflecting evolutionary distance and allows comparison between taxa [91]. These distinct characteristics of the the 16S rRNA gene make it an ideal target for amplicon sequencing and the polymorphisms specific to a bacterial lineage can be used as a molecular 'fingerprint' to profile the taxonomic composition of a microbial community.

Due to the small length of the 16S gene, sequencing efforts are inexpensive; this has resulted in numerous 16S sequence deposits in public databases. One such database called Greengenes [92], contains over one million 16S sequences; BLAST [93] queries can be performed on these to annotate unknown 16S sequences. Despite the many positive characteristics of the 16S rRNA gene, amplicon sequencing is unlikely to capture the entirety of bacterial diversity as the primers designed are based upon previously isolated microbes which may not be characteristic of all prokaryotes.

2.1.1.2 Taxonomical characterisation

Upon obtaining 16S reads, the sequences are assigned to a taxonomy class for further analysis of bacterial composition. Several tools exist for taxonomic characterisation, although the most commonly used methods include Mothur [94] and Quantitative Insights Into Microbial Ecology (QIIME) [95].

The task of annotating sequences with taxonomic information is performed by clustering sequences into Operational Taxonomical Units (OTUs). An OTU represents a group of highly similar sequences with small phylogenetic distance intended to approximate bacterial species. Sequences are binned at different thresholds of sequence similarity to achieve

different levels of taxonomic resolution. Binning at 97% similarity is thought to approximate differences at the species level, however in practice, similarity thresholds are arbitrary and are dependent on the research question and values of 95% to 99% are typically used [87].

Clustering of sequences and the picking of OTUs is an important step which impacts upon on downstream analysis. In the QIIME pipeline, OTUs can be defined by three main strategies: De novo, open reference, or closed reference [95]. 'Closed reference' OTU picking aligns reads to a reference database resulting in a high quality phylogenetic tree, however, reads that do not match to a known sequence are removed thereby discarding potentially novel sequences. 'De novo' clusters reads based on sequence similarity independent of a reference database so that all reads are retained, however, the taxonomy of de novo OTUs are unknown. 'Open reference' is a hybrid approach which clusters sequences using a reference database like that in closed reference, however, reads which do not match in the database are clustered using the de novo approach. Open reference OTU picking is often used as it retains the ability to assign informative taxonomic annotations as well as detecting novel diversity. Once OTUs are picked, a phylogenetic tree is constructed by a multiple sequence alignment of OTU sequences with a tree building algorithm such as FastTree [96]. Upon completion of OTU picking, the resultant taxonomical classifications are summarised in a count matrix $\mathrm{C}(c_{i,j})$ where $c_{i,j}$ is the number of reads for an OTU i in a sample j which is used for further downstream analysis.

2.1.2 Analytical methods for Microbiome data

This section will consider the analytical methods applied to microbiome data including the preparation of raw data, analysis of community composition via α and β-diversity measures, ordination, differential abundance and co-occurrence analysis.

2.1.2.1 Data normalisation

Before comparisons between pre-defined groups, a normalisation step must be performed to account for the technical effects of sequencing and to allow comparisons between samples. Samples are often sequenced to uneven depths which range in orders of magnitude [97]. Differences in library sizes across groups can result in the clustering of samples by sequencing depth [98] instead of true biological variation. Further, if left uncorrected, unequalised library sizes can inflate the rate of false positives and any observed differences in

OTU abundance could be due to technical and not biological effects [98].

One of the most popular methods for normalising microbiota count data data is rarefying. This procedure equalises the library sizes across samples by randomly subsampling sequences to a fixed depth. A minimum library size is chosen and all samples below this threshold are removed from the dataset. This threshold is selected in order to remove only a small percentage of under-sampled communities. The remaining samples are then re-sampled to the common minimum library size. Rarefying has been utilised in many publications and is part of the standard QIIME pipeline [95], however, it has been criticised and described as 'statistically inadmissible' [97] as the procedure throws away part of available data and inflates variance due to random subsampling.

Other normalisation methods derive a normalisation factor f_j representing the true library size of sample j. The normalisation factor is used to globally adjust the counts in a sample allowing for comparison across libraries [99]. The most widely applied method is total sum scaling (TSS) which uses the total library size of each sample as the scaling factor f_j. TSS does not introduce additional noise by randomly subsampling and transforms counts into proportions on a scale of 0-1 which are often called 'relative abundances'. Other approaches have been developed specifically for microbiome data. These include cumulative sum scaling (CSS) [100] in which the normalisation factor is derived from the total sum of reads up to a quantile which is estimated from the data. The underlying concept is that a sequencer will preferentially sample abundant sequences, and calculating scaling factors from the total library size will penalise sequences of lower abundance. CSS thus attempts to find an appropriate quantile and normalisation factor such that less influence is afforded to highly abundant species.

Given the similarities between 16S and RNAseq data, recent studies have explored the utility of methods initially designed for RNAseq data [97, 98]. These include trimmed mean of M-values (TMM) [101] which is a method included as part of the *EdgeR* package [102]. TMM normalisation uses the total library size, however, a normalisation factor is first calculated to correct the library size for each sample. TMM works under the assumption that most features are not differentially abundant and the normalisation factors are calculated after removal of differentially abundant taxa [99]. This means that the scaling factors are derived on the 'non-differentially abundant' component of the microbiota

and places less influence on highly abundant sequences. Normalisation of microbiome data is an intense area of research and debate, however, there is no agreed consensus [97, 98, 103].

2.1.2.2 Alpha diversity

It is common to describe a microbial community by means of α-diversity which profiles the species composition *within* a sample. Two main components of α-diversity are commonly used in microbiota studies: richness and evenness.

Species richness refers to the absolute number of unique species present in an environment. The simplest measure of species richness is to count the number of individual species in a sample, however, this approach can be sensitive to technical issues such as undersampling and minor differences in DNA concentration. The Chao1 richness estimator [104] assumes that if a sample has many rare species, i.e., singletons, then it is likely that there are more species which were not detected. Chao1 therefore estimates the true species richness whilst taking into account under-sampling based upon the number of singletons and doubletons in a sample. The second measure, evenness reflects the homogeneity of species abundances in a community. For example, a sample consisting of approximately equal species proportions reflects high evenness, whereas a sample with an extremely dominant species has low evenness.

To quantify species diversity, hybrid measures are often used to summarise both species richness and the evenness of the community into one measure. One such popular approach is the Shannon index [105] which calculates the sum of species proportions in a community whilst accounting for the total number of species, S, and is calculated as:

$$\mathbf{H}' = -\sum_{i=1}^{S} p_i \ln p_i \tag{2.1}$$

where p_i is the proportion of species i. p_i is estimated as $p_i = n_i/N$ where n_i is the count of species i, and N is the total counts for all microbes identified in a community. The Shannon index therefore increases with greater richness and evenness.

As alpha diversity is calculated for single community (sample) it is often used to compare diversity between predefined groups such as disease status or body sites. In clinical microbiota studies, it is generally considered that higher diversity is a beneficial trait and

protects against offending microbes whilst states of lower diversity can indicate dysbiosis and a loss of homeostatic equilibrium [106, 87].

2.1.2.3 Beta diversity

Alpha diversity describes the composition at a single site, however, a major objective of many microbiome analyses is to describe differences in community composition between sites [14]. β-diversity is used to compare the compostion between communities and is calculated as a measure of distance or dissimilarity between two samples. In most cases, more than two samples are to be compared, therefore, β-diversity is calculated between every pair of samples to generate a square distance or dissimilarity matrix.

For studies of the microbiota, there are two main classes of β-diversity measure. The first type consider only species abundance patterns, known as taxon based, whereas the second also take into account the phylogenetic distance between species [107]. Common measures of non-phylogenetic β-diversity consist of Bray Curtis (BC) [108] dissimilarity and the Jaccard index. Bray Curtis measures pairwise dissimilarity between sites using the shared abundances of species as well as the total abundance at each site. The calculation is robust to zero counts and as microbiome data is characteristic of many zeros, BC is one of the most popular metrics for analysis of β diversity in microbiome studies [103].

Measures such as the Jaccard and BC treat species as independent features, however, newly developed measures also consider the phylogenetic relationship between species. For example, two species of the *Lactobacillus* genus are more genetically similar than two species of the *Lactobacillus* and *Peptostreptococcus* genus. Phylogenetic β-diversity measures work under the assumption that closely related species share many of the same functions representing redundancy, and should be down-weighted compared to shared counts between divergent species. Distance measures of this class have been designed specifically for microbiome studies such as the UniFrac distance [107]. Weighted UniFrac distance takes into account both abundance and phylogenetic distance whereas unweighted UniFrac only considers branch length of the phylogenetic tree as well as presence-absence patterns.

2.1.2.4 Ordination

Calculation of beta diversity results in a distance matrix describing the pairwise dissimilarity between samples. The resultant distance distance matrix is high-dimensional and thus cannot be visualised. Ordination of the distance matrix projects the high-dimensional sample distances into a low-dimensional 'ordination' space such that the relatedness between samples can be observed on a small number of axes (typically two to three). The objective of ordination algorithms is to represent sample distances in a low dimensional space whilst preserving the true distances as closely as possible. Whilst there are many available methods for ordination, two of the most popular are principal co-ordinates analysis (PCoA) and non-metric multidimensional scaling (NMDS). PCoA is performed by singular value decomposition of the distance matrix and finds the directions of greatest variability between sites and projects them into a low dimensional space. PCoA assumes a linear relationship between sites, however, if non-linear trends are present PCoA may fail to accurately represent the true distance resulting in an 'arch effect' [109]. NMDS, on the other hand, attempts to find the best position of samples in a pre-specified number of dimensions using an iterative procedure whilst minimizing 'stress' [109]. Stress is a measure of the distance between points in ordination space compared to the true distances, thus a low stress ordination solution closely represents the true distances. NMDS uses rank distances meaning that it is robust to non-linearity and can be used to ordinate a wide range of dissimilarity metrics.

2.1.2.5 Differential abundance analysis

A major objective of microbiome analysis is to identify a change in abundance for a specific microbe in response to disease and or treatments. Such analysis is important to identify species which may be lost in disease and may have a protective effect, or those which increase and could be pathogenic. The term 'differential abundance analysis' refers to the statistical challenge of profiling systematic changes in OTU abundance across pre-defined groups.

16S data is usually non-normally distributed, therefore, common approaches are to apply non-parametric tests such as the Wilcoxon-rank sum in the case of two classes, or Kruskal-Wallis for multiple groups after normalisation of the OTU table. These statistical tests have been used to identify differentially abundant OTUs in many studies [85, 64, 110]. Other methods have been specifically designed for 16S data including LEfSe [111], which

is a tool based upon Kruskal-wallis and Wilcoxon tests coupled with linear discriminate analysis.

Non-parametric tests do not assume a distribution and are an appropriate test for differential abundance analysis, however, they have been shown to suffer from limited sensitivity [98] and do not have the capability to control for environmental factors or covariates. As a result, there has been an increasing preference for the application of parametric tests to studies of the microbiota due to their increased power, particularly in the detection of low abundance OTUs, and their ability to quantify effect size. One such example is Metastats [112] which uses a two sample t-test with permutation testing to estimate significance with respect to the non-normal distribution.

Recent research has focused on the application of parametric models intended for RNAseq analysis to studies of the microbiota; these are based upon the negative binomial distribution (NB) [97, 98]. Historically, RNA seq counts were described with a Poisson distribution [113], however, when taking into account biological replicates, the variance in read counts is greater than the Poisson distribution can explain [97]. This phenomenon, known as overdispersion, means that the Poisson distribution tends to underestimate the variance resulting in an inflated rate of false positives [97]. The negative binomial distribution has an extra parameter which can take into account overdispersion allowing for a better fit of biological replicates. Examples of such methods incorporating the negative binomial distribution include DEseq2 [114] and EdgeR [102], both of which have been shown to have higher sensitivity than non-parametric statistics in recent comparative analysis [97, 98]. These models use raw read counts to estimate their own normalisation factors - TMM in edgeR [101], and RLE in DESeq2, which are used to adjust for uneven library sizes. Despite promising results, methods based upon the NB distribution do not always perform well with excessive zeros [98]. Other approaches include metagenomeSeq [115] which is based upon zero-inflated Gaussian (ZIG) mixture models that assume excessive zero counts are not a result of absence but are due to undersampling. As with normalisation strategies, there is a considerable debate as to which is the best approach for differential abundance analysis and no consensus has been reached [98].

2.1.2.6 Linear models for microbiome analysis

The application of linear models for differential abundance, as well as for omics' integration has gained in popularity due to its ability to control for multiple potentially confounding variables. Extraneous sources of variation including sex [40, 41], age [40], body site [33, 38], diet [18, 116, 19], and ethnicity [117] have all been shown to impact upon OTU counts and multiple regression allows for estimation of the linear relationship between the response and predictor of interest independent of all other factors included in the model.

The most well known package implementation for microbiome data is MaAsLin [5, 118] which incorporates linear models with an arcsine square-root variance stabilising transformation to account for the heteroscedasticity of OTU proportions. For each OTU, a linear model is fit of the form:

$$arcsin(\sqrt{Y}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \tag{2.2}$$

where Y is the proportion of an OTU, β_0 is the intercept term, X_p are the metadata to be tested, and β_1 is the coefficient representing the marginal change in $arcsin(\sqrt{Y})$ for a one unit change in β_1 whilst all other β_p are held constant.

The ability of linear models to quantify an association independent of other factors has resulted in a wide range of applications, not only including differential abundance analysis for disease association [6, 5], but also for host-microbiota and metadata-microbiota interactions [119, 120, 29, 121].

2.1.2.7 Co-occurrence analysis

Differential abundance analysis is a reductionist approach which considers taxa as independent entities, however, in reality, taxa interact with each other in intricate and complex ecological systems. Whilst many conditions, particularly those with a genetic basis are associated with a single causal mutation, it also clear that many phenotypes are complex and are linked to multiple factors. It is therefore just as likely that a host phenotype is not just driven by a single microbe, but by a group or community of pathogenic microbiota. The underlying principal of co-occurrence analysis is that when pairs of species tend to be present in the same communities often, there is likely to be an association between them. Co-occurrence analysis is thus the identification of these interactions and the patterns of

species abundances across samples or conditions.

Methods based upon presence and absence patterns have been successfully used to identify co-occurrence patterns in studies of the microbiota. These include studies which have used the checkerboard score [19], as well as the dice index [122], however, co-occurrence patterns can also be represented as a network in which OTUs are modelled as nodes and associations between taxa as edges. The underlying concept is that a community of highly connected OTUs may be functionally related providing insight into the underlying ecological system. Whilst correlation based approaches using Pearson or Spearman metrics have been used to reconstruct ecological interaction networks [14, 123], they have been shown to result in spurious correlations as relative abundance data is compositional [124, 125]. This means that if the relative abundance of one species goes up then another must come down. To account for this, methods specifically developed for microbiome analysis can be applied such as SparCC [124] which estimates the correlation using the variances of the log-ratios between taxa.

2.2 The Transcriptome

According to the central dogma, genes encoded within the DNA are transcribed into messenger RNA (mRNA) by a process called transcription. Within a specific environment, the totality of these transcribed DNA sequences is called the transcriptome. Whilst the genome is static, the transcriptome is dynamic representing the components of the genome which are actively transcribed in a tissue at a given time point. Microarrays, which have emerged from the human genome project, have transformed mRNA analysis by allowing simultaneous profiling of thousands of genes. Microarrays allowed researchers who had become accustomed to focusing on small subset of genes to consider global shifts in gene expression enabling detection of new phenotype responsive genes. Whilst initially microarrays were used to screen for associations amongst many individual genes, now they can be used to understand the role of genes at the systems level.

High throughput omics' technologies are constantly evolving and few technologies have contributed as much to our understanding of complex disease. Gene signatures have now been unearthed for many pathologies including cancers as well as response to treatment and offer promise for future development of therapeutics.

2.2.1 Methods to analyse the transcriptome

2.2.1.1 Microarrays

Hybridisation-based methods offer an inexpensive way to obtain transcriptomic information from biological samples. A DNA microarray is a surface on which a high density of probes are fixed at defined locations, each of which represents a specific gene encoded on the genome [126]. The probes are constructed from millions of short 25 bp single stranded DNA molecules known as oligonucleotides which act as biosensors for complementary RNA (cRNA) targets. The cRNA targets are extracted from a biological sample of interest and then bind to the probes in a process called hybridisation. Quantification of the amount of target RNA binding at a specific probe can be measured by labelling targets with a fluorescent marker and scanning with a high resolution epifluorescent microscope [126]. The fluorescent intensity level of each probe is quantified to obtain genome wide relative measures of gene expression. Despite suggestions that microarrays have reached their technical limit, and are to be replaced by more modern technologies such as RNASeq which are not limited by a defined probeset [127], microarrays remain a powerful tool for obtaining insight into the transcriptional architecture of biological systems.

2.2.2 Analytical methods for transcriptomics data

This section will consider analytical methods applied to transcriptome data including the preparation of raw data, identification of differentially expressed genes, methods for functional analysis and dimensionality reduction techniques.

2.2.2.1 Data normalisation

Technical errors introduced by 'cross-hybridisation', a phenomenon when target RNA hybridises non-specifically, as well as instrumental and biochemical factors all contribute to the noise associated with microarray experiments. The changes in raw expression values across arrays are thus composed of both technical and biological effects. As a consequence, it is important to normalise intensity values to minimise technical variation, and to allow for meaningful comparison between arrays.

Early methods of normalisation scaled the global intensity values in a sample by a constant such as the mean or median intensity across arrays [128]. Others realised that the expression of 'housekeeping genes' were constitutively expressed and used the expression of these

genes as reference for normalisation. These genes were considered to vary little between conditions, however, this was suboptimal as the high expression of housekeeping genes is not representative of most genes, and their expression was not as stable as initially thought [129]. Modern approaches for normalisation of microarrays include the MicroArray Suite 5.0 (MAS5) [130] algorithm which corrects for background noise using mismatch probes (mm) and then a normalisation step is performed independently for each array based upon a robust average of background corrected intensities. As each array is normalised independent of other arrays, MAS5 is not dependent on the sample size or quality of individual arrays and is particularly useful in circumstances where additional samples may be added at later dates [131].

An alternative approach is used in the robust multiarray averaging (RMA) method [132] which does not use mismatch probe information. Instead, RMA estimates the true signal using a convolution model under the assumption that probe intensities are composed of a signal and a common background noise component. RMA then uses quantile normalisation to force the background corrected intensity distribution of all arrays to be equivalent. Modern Affymetrix chips such as the HuGene 2.1st do not include mismatch probes which limits the application of MAS5 and in such circumstances normalisation with RMA is preferred.

2.2.2.2 Differential analysis

Transcriptomics data enables the identification of molecular signatures associated to a pathology or treatment. The statistical task of identifying group responsive genes comes under the category of differential expression analysis (DA), where the objective is to identify differentially expressed genes (DEGs). Identification of DEGs is important as signatures of up-regulated and down-regulated transcripts can provide a basis for understanding pathological mechanisms.

Historically, DEGs were identified by calculating the fold change in mean expression values between predefined groups. Whilst an intuitive approach, fold changes do not consider sample variances and can fluctuate greatly if the denominator is small, as is often the case with lowly expressed genes. Further analysis showed that calling differential expression based solely on a fold change criterion resulted in unacceptable levels of false positives [133, 134]. To overcome problems with the fold change criterion, statistical tests were used

as they take into account sample variances and also allow the calculation of p values. Early studies used t-tests, however, the t-test suffers from low power with small sample sizes. Problems with t-tests were also compounded by high false positives rates amongst lowly expressed genes [135]. These genes can have low variances resulting in a large t statistic even when the difference in means is small.

To overcome the shortcomings of t tests, more complex methods were developed such as as Significance Analysis of Microarrays (SAM) [136], however, the most popular method used today is linear models for microarray data (limma) [137, 138]. The principle underlying of limma is to 'borrow information' across all genes on the array and combine the gene specific variances with a population estimate in an approach known as 'shrinkage'. This ensures that genes with low variances are not falsely differentially expressed in the absence of an acceptable difference in means. Whilst both SAM and limma are used in modern microarray analyses, a recent study concluded that SAM had weak performance and was outperformed by limma [139]. Many modern transcriptome analyses today combine both p values and fold change criteria to ensure that associations are significant, whilst focusing on the most biologically relevant changes [58, 140].

2.2.2.3 Multiple testing

Gene expression data consists of many features which are often orders of magnitude greater than the number of samples measured. For a study of 10,000 genes, differential analysis at an α level of 0.05 would result in 500 DEGs just by chance alone. This, known as the 'multiple testing problem', an issue which plagues hypothesis testing in high throughput analyses. One way to reduce false positives is to control for the family-wise error rate (FWER) using Bonferroni correction. This method aims to reduce the experiment-wide probability of making a false positive to α [133] meaning that for the same study, a p value of 0.05 / 10,000 = 5e-06 would be required for a gene to be called differentially expressed. This is considered to be too conservative for many biological analyses which often have small sample sizes. Instead of controlling the experiment wide error, false discovery rate (FDR) correction aims to control the proportion of false positives amongst the genes considered to be differentially expressed. Methods which control the (FDR) such as the as the Benjamini Hochberg method [141] offer greater sensitivity for discovery at the cost of increased type-I error.

2.2.2.4 Functional analysis

The functional properties of genes which have been characterised through experimentally or computationally derived methods are stored in rich publicly accessible repositories such as Gene Ontology (GO) [142]. As well as ontologies, several databases hold mechanistic information about the physical interactions between biomolecules known as biochemical pathways. Pathways are often curated using the available literature describing the mechanisms underlying particular biological responses. Databases of this kind include Ingenuity Pathway Analysis (IPA) [143], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [144], and REACTOME [145].

For a list of differentially expressed genes, it is often of interest to determine if its constituents are statistically associated with a functional process or pathway. Enrichment analysis, also known as over-representation analysis (ORA), integrates experimentally derived gene lists with ontology and pathway databases. The overall objective is to determine if an experimentally determined list has a higher gene-specific overlap with a pathway than would be is expected by chance. Tests to perform ORA use contingency tables representing the number of genes within a pathway, and the number of genes which were found to be significantly differentially expressed. A statistical test such as the hypergeometric or Fisher's exact test is then used to determine if genes within a pathway are over-represented with respect to the total numbers of genes in the pathway, and the total number of genes under investigation.

Over representation analysis considers genes as equally weighted features which are dependent on a hard threshold. Whilst a cut off of $p < 0.05$ is an accepted criterion for differential expression, this threshold includes genes $p = 0.049$, and discards genes $p = 0.051$ which could potentially hinder the detection of biologically relevant pathways [146]. Furthermore, no information regarding the magnitude of differential expression is considered and the reductionist approach of considering genes as individual entities is not an accurate representation of living systems. Other strategies for pathway enrichment such as Gene Set Enrichment Analysis (GSEA) [147] address some of these issues by considering a ranked list of t statistics or fold changes instead of a subset of the most DE genes. The question is then asked whether genes within a given pathway are found in a greater quantity towards the top (indicating upregulation) or towards the bottom of the DEG list by calculating a running-sum statistic. As an enrichment score is computed for each pathway

dependent on the expression changes of the whole set of pathway members, subtle changes to the gene set in concert contribute a greater enrichment than changes in a small number of genes which may be more biologically relevant.

2.2.2.5 Dimensionality reduction

High throughput omics' experiments suffer from the 'small n large p' problem in which the number of variables measured greatly exceeds the number of observations. This is problematic as the number of pairwise tests required to perform a comprehensive unbiased analysis is large and exhaustive pairwise testing inflates Type I error unless stringent multiple testing correction is undertaken. The main problem with FWER or FDR correction is the subsequent loss of power to detect true differences, however, one way to overcome this is to reduce the number of hypothesis tests performed with a suitable scheme of dimensionality reduction.

The most popular method for dimensionality reduction of transcriptomics data is principal components analysis (PCA). PCA is a linear function which summarises the variables contained in a high dimensional dataset into a set of orthogonal factors called principal components. PCA is achieved by performing eigen-decomposition of the covariance (or correlation) matrix where the eigenvectors correspond to the principal components, and the eigenvalues correspond to the variance explained by the corresponding eigenvector. In the context of transcriptomics data the principal components (PCs) are linear combinations of genes representing the maximally variable directions in multidimensional space [148]. In multidimensional space, the first PC is the direction which explains the most amount of variance, followed by the second component which is the next most variable direction which is also orthogonal to the first PC. In an example proposed by Ringer et al. [148], 100% of the variance of 8,534 genes was explained by only 104 components. PCA is often used for visualisation of samples projected onto a small number of components to observe global trends and is also used for data integration [29].

Others have expanded the utility of PCA by incorporating prior knowledge of gene sets from pathway databases. The underlying idea is to reduce the hypothesis testing space from tens of thousands of genes to hundreds of pathways by computing a 'pathway activity' score which can then be used for differential analysis. One such example is PLAGE [149] which performs singular value decomposition of the pathway specific expression matrix.

The major limitation with biochemical pathways is that they are usually a static representation of healthy states [146], which are not characteristic of the dynamic processes which underlie disease. For this reason, dimensionality reduction techniques which infer the 'set' of genes directly from the data may be preferred. One such method is Weighted Gene Co-expression Network Analysis (WGCNA) [150, 151] which reconstructs gene-gene co-expression networks and performs community detection in order to identify modules of highly co-expressed genes. Methods such as this assume that individual genes act in groups of highly connected genes forming modules. Module members have similar expression patterns are thus considered to be functionally related due to the principle of 'guilt by association'. Singular value decomposition of the module-specific expression matrix, which is equivalent to the first principal component, is used to calculate the module eigengene which represents a pattern of module-centric gene expression. The module eigengene can then be used to detect differences between groups, or for relating module expression with clinical traits reducing the hypothesis space from tens of thousands of genes to tens of modules.

2.3 The MAARS cohort

The datasets analysed in this thesis were generated by the Microbes in Allergy and Autoimmunity Related to the Skin (MAARS) consortium. A description of the protocol used has been prepared by MAARS consortium members and is quoted in Sections 2.3.1, 2.3.2 and 2.3.3.

2.3.1 MAARS Subject recruitment and sampling

"To evaluate differences in the cutaneous microbial colonization and transcriptional profile in atopy- and autoimmune-type skin diseases, adult patients (18-70 years) with mild-to-severe chronic AD (SCORAD score > 25, n=88) and plaque-type PSO (PASI score >7, n=129) as well as healthy volunteers (n=117) were recruited from three Depts. of Dermatology, at University Hospitals located in Duesseldorf (HHU, Germany), London (KINGS, Great Britain) and Helsinki (UH, Finland). Each subject underwent a physical examination by a dermatologist and the medical history was recorded. The diagnoses were made by a dermatologist based on clinical presentation, personal history, laboratory findings and the criteria of Hanifin and Rajka [152]. The exclusion criteria included concomitant

autoimmune diseases (e.g. rheumatoid arthritis, diabetes, alopecia areata, etc.) the use of systemic antibiotics within 2 weeks and systemic immunosuppressive therapy or phototherapy or systemic biologic agents within the previous 12 weeks prior to screening. Before skin sampling, the biopsy sites were left untreated for at least 2 weeks and cleansing with only the non-antibacterial Dove soap was allowed and washing was avoided for 24 hours prior to sampling. The patients or healthy volunteers who did not match these clinical exclusion criteria were removed from the study. The following biological samples were then obtained and submitted to analysis: 1) microbiome samples from upper/lower back, posterior thigh or buttocks (PSO, AD, healthy volunteers) with no prior cleaning or preparation of the skin surface using sterile gloves to prevent cross-contamination, were obtained placing a sterile ring (2.5 cm diameter) onto the appropriate skin area, 1.5 ml PBS was supplemented into the ring and the area sampled scraping a glass rod in a circular motion 10 times to the left and to the right. Subsequently, the microbiome-enriched PBS was harvested and stored. In addition, mock samples containing only PBS were collected at each sampling time in order to assess contamination. 2) 6 mm punch biopsies from skin at the microbiome sites were taken in local anaesthesia. Subsequently, samples were stored in RNAlater (Sigma-Aldrich) and subjected to further analyses. The study was approved by the appropriate local Institutional Review Boards (UH, Dnro 91/13/03/00/2011; HHU, 3647/2011; KINGS, 11/H0802/6) and all subjects provided written informed consent before participation." (MAARS consortium. *Unpublished*, June 2016)

2.3.2 MAARS Microbiome processing

"**DNA extraction.** DNA was extracted from the clinical swab and mock samples using Qiagens Pathogen Lysis Tubes and the QIAamp UCP Pathogen Mini Kit (Cat.No: 19092) according to manufacturers instructions. In brief, sample pellets were resuspended in 500 l Buffer ATL and vortexed for 10 min at maximum speed using Pathogen Lysis Tubes containing glass beads. The samples were transferred to fresh Beckman tubes and incubated in 16.5 mg/ml lysozyme (Sigma) for 30 min at 37C. 50 l proteinase K were added and the samples were then incubated for 10 min at 56C. Addition of 250 l of Buffer APL2 was followed by incubation at 70C for 10 min. 10 l RNA-grade glycogen (20mg/ml, Thermo Scientific) were added to maximize DNA recovery. Ethanol was added to a final concentration of 25%. DNA was extracted and washed using spin columns, and subsequently eluted in 50 l of Buffer AVE. **16S rRNA gene amplification and preparation for sequencing.** 2.5 l template were amplified in RT-PCR GradeWater (Life technologies), 3% DMSO,

with 1x PCR HF buffer using Phusion Hot start II DNA polymerase, 200 M dNTPs (all Thermo Scientific), and 500 nM custom primers (Eurofins MWG Operon). One universal forward primer (341f 5-CCTACGGGNGGCWGCAG with adaptor B, Lib-L) was paired with one of 104 barcoded reverse primers (805r 5-GACTACHVGGGTATCTAATCC with adaptor A, Lib-L). Each barcode consisted of seven nucleotides, contained no homopolymers, and a pair of barcodes differed in at least 2 positions. Each PCR was run in triplicates and the PCR products from each sample were pooled. A negative control PCR reaction lacking template was included for all primer pairs in each run. The PCR was run for 30 cycles. The PCR products were purified from the reaction using Dynabeads MyOne Carboxylic Acid (life technologies, Cat.No: 35401) and TruSeq precipitation buffer (16% PEG-6000, 1.5 M NaCl) on the Magnatrix 1200 (LBH Advanced Bioservices AB, Sweden). The purity of the amplicons was visualized on the Agilent 2100 BioAnalyzer using High sensitivity DNA chips and reagents (Agilent Technologies, Cat.No: 5067-4626) according to manufacturers instructions. DNA concentrations were measured by real-time PCR (KAPA Library Quantification Kits For Roche 454 GS Titanium platform, Cat.No: KK4821 and BioRad CFX96 Touch Real-Time PCR Detection System; C1000 Thermal cycler) according to manufacturers instructions with samples diluted 1:500, 1:1000, and 1:2000 in 10 mM Tris-HCl, pH 8.0. Extension time was 90 sec. Finally, the samples were adjusted to 1.0*108 DNA molecules for each sample before pooling 50-60 samples per 454 sequencing run. **454 amplicon sequencing.** EmulsionPCR was performed on the amplicon library using a large volume emPCR (Lib-L, v2 reagent kit) according to the manufacturers amplicon protocols and pyrosequenced (one way read direction) on a Genome Sequencer FLX-Titanium instrument (Roche/454 Life Sciences) at Science For Life Laboratory (SciLifeLab) Stockholm. Each library was sequenced in both regions of a two region gasketed 7075 mm Titanium PicoTiterPlate, and base calling was performed with the on-instrument amplicon filter settings. Samples containing only water were sequenced in order to assess contamination during the sequencing process. **Demultiplexing and preprocessing of 454 reads.** All sequence reads were assigned to their samples using the unique sample barcodes. Raw sequence reads were analyzed with AmpliconNoise version 1.25 [153] to remove 454 sequencing and PCR artifacts and PerseusD from the same program package to remove PCR chimaeras, using default parameter values. The output from each sample was further processed in QIIME (Quantitative Insights Into Microbial Ecology) version 1.8.0 [95] if the number of processed high quality reads exceeded 3000 per

sample. Otherwise, the sample was resequenced. **OTU clustering and taxonomy assignment**. The preprocessed dataset comprised of a total of 3,357,091 high quality reads. The following analysis steps were performed using QIIME version 1.8.0. [95] OTUs were picked at 99.3 % identity using the `pick_open_reference_otus.py` command and uclust 1.2.22q [154]. Taxonomy was assigned using blast-2.2.22 [93]. The reference data files used for both OTU clustering and taxonomy assignment were downloaded from the Greengenes Database Consortium [92]). As AmpliconNoise did not perform very well in identifying chimeric sequences in our dataset ChimeraSlayer [155] was applied here within the QIIME pipeline and identified chimeric sequences were removed from the OTU table and the phylogenetic tree. Three samples of poor quality were removed from the OTU table. Abundances were normalized using the Trimmed Mean of M-values method (TMM), implemented in the edgeR Bioconductor package [102]." (MAARS consortium. *Unpublished*, June 2016)

2.3.3 MAARS Transcriptome processing

"The tissue samples were stored in RNAlater and total RNA was extracted from the tissue samples using the RNeasy Fibrous Tissue Mini kit (Qiagen). Tissue samples were homogenized using the FastPrep-24 instrument (Nordic Biolabs AB), and RNA was extracted according to the manufacturers instructions. The yield and purity of RNA in the samples were controlled using a Nanodrop spectrophotometer and Qubit fluorometer to verify absence of inhibitors (R260/280: 2.1; R260/230nm: 1.3). RNA integrity was quantified by electrophoresis and performed using Agilent dedicated Lab-on-chip (RNA6000 Nano and Pico kits). RNA Integrity numbers and 28S/18S ratio averages were respectively 8.6 and 2. 100ng of total RNA were amplified according to Affymetrix protocols (Affymetrix GeneChip Whole Transcript (WT) Expression Arrays). Based on expertise of Institut Curie genomic platform, MAQC A RNA samples (Universal RNA, Stratagene, P/N: 740000) were implemented to series of RNA amplification in order to monitor target preparation. In practice, series of 47 RNA (from healthy volunteers, and patients) and 1 Universal RNA were amplified, monitored and labelled. During synthesis steps, purified molecules were quantified using a multichannel Nanodrop (ND8000, Thermo) to normalize amount of molecules used for DNA synthesis (10g) and hybridization (5.5g). Molecules were also controlled on high throughput electrophoresis (QIAxcel DNA, Qiagen) in order to monitor size of complementary RNAs (average: 500nt), and fragmented DNA (average: 50nt), to ensure quality of targets and hybridization of microarrays. Series of 96 targets

were hybridized onto Affymetrix Gene ST 2.1 96 array plates, including in total two Universal RNA, using an Affymetrix Genetitan MC system. Quality of raw data and normalized data were monitored to control dynamics of the measurements, across series of synthesis, and series of hybridization using bacterial spike in controls added to total RNA, and using Universal RNA. An automated quality control pipeline based on the arrayQualityMetrics method [156] was used to capture quality failures in microarray data. Data were then normalized using the Robust Multi-array Average (RMA) [132] approach implemented in the affy Bioconductor package [157]." (MAARS consortium. *Unpublished*, June 2016)

Chapter 3

The skin microbiome in homeostasis and disease

3.1 Introduction

The skin is a primary interface and acts as a first line of defence against invading pathogens, preventing infection and exposure to harmful compounds. The vast resident population of microrganisms which cohabit the human skin exist in symbiosis and have convolved with their human hosts [18]. It is now accepted that some of these commensal species are beneficial and help to maintain a state of homeostatic equilibrium on the skin, promoting host health and preventing infection of pathogenic species [31, 22].

Studies of the healthy skin microbiota indicate that community composition is dominated by 4 main phyla consisting of *Firmicutes*, *Proteobacteria*, *Actinobacteria* and to a lesser extent, *Bacteroidetes* [38, 31, 158, 41]. A diverse and healthy microbiota is thought to promote host immunity [159] and as the skin contains a plethora of immune cells [160], it is possible that host-microbiota interplay is associated with immune system training [161]. In the event of dysbiosis, which reflects a loss of diversity, this balance is disrupted which could lead to inappropriate immune response and inflammation. Currently it is unknown whether dysbiosis of the commensal bacteria is either a cause or effect of disease [31].

Host factors such as body site, age, gender [40, 31, 41, 38] and geographic location [117] have been shown to be associated with community composition. Furthermore, studies have shown that healthy skin harbours a complex and diverse ecosystem and is one of the organs

with the most heterogenous microbiota [1, 39].

Atopic dermatitis is an allergic disease characterised by increased IgE levels, dysbalance of the Th1/Th2 axis, and abundance of Th2 cytokines [44]. Barrier dysfunction in AD [51] may increase permeability to microbial antigens [52] resulting in improper immune cell polarisation; thus, is plausible that dysbiosis could result in inappropriate immune activity. Studies of resident bacteria have shown high abundances of *Staphylococcus aureus* on both the lesions and uninvolved skin of AD patients [62]. As with other diseases, AD is characteristic of a dramatic loss in species diversity [64]. Furthermore, *Staphylococcus aureus* has been shown to be of higher abundance during an inflamed state compared to the pre and post-flare phases [64]. Moreover, it has been proposed that *S. aureus* abundance may increase from baseline during the pre-flare phase.

Psoriasis is characterised by extreme quantities of antimicrobial peptides (AMPs) and increased infiltration of Th1, Th17 and innate immune cells [73]. Previous analysis of the psoriatic microbiota have presented variable and inconsistent results. An early study [84] identified an expansion of *Firmicutes* compared to healthy skin, and a reduction of *Proteobacteria* and *Actinobacteria*. Furthermore, a strong under-representation of *Propionibacterium acnes* was found on lesional skin which the authors suggested may play a beneficial role, or may be displaced by more dominant species. More recent results showed that psoriatic skin was characteristic of reduced species richness and increased abundance of four genera including *Corynebacterium*, *Propionibacterium*, *Staphylococcus* and *Streptococcus* [85]. Contrasting results from a study of the microbiota isolated from skin biopsies identified reductions in *Actinobacteria* as well as increases in *Proteobacteria* [86].

Given the inconsistencies of previous analyses, further study of the disease associated and inflamed microbiome could provide important insights into understanding the role of microbiota in health and disease. Previous analysis of the inflammatory microbiome consisted of small sample sizes, and no other study has directly compared the microbial composition of AD and PSO in a single analysis. Here using the MAARS cohort, this chapter presents an exploratory analysis into the largest microbiome cohort to date consisting of 87 patients with Atopic dermatitis, 128 patients with Psoriasis and 117 healthy volunteers to highlight core differences with respect community composition in disease.

Parts of this chapter are extensions to preliminary unpublished analysis performed in collaboration with MAARS consortium members. In these cases, the analyses have been extended and modified to address my own research questions. Species diversity analysis, specifically the comparisons of α-diversity in **Figure 3.3 B**, O_2 tolerance in **Figures 3.3 E-G**, and analysis of β-diversity in **Figure 3.4**, build upon previous works by Stefanie Prast-Neilsen, Björn Anderrson and Marine Jeanmougin. These analyses have been extended to assess the association with disease severity, as well as incorporation of higher order taxonomy and non-lesional disease cohorts.

Differential abundance of OTUs build upon collaborative unpublished analyses by Stefanie Prast-Nielsen, Mauricio Barrientos-Somarribas and Björn Anderrson. The original analysis compared OTU relative abundances between healthy volunteers, AD-lesional and PSO-lesional samples and then significant OTUs were evaluated and corrected post-hoc for associations with non-clinical factors. The work presented in this chapter was extended to include higher order taxonomy from the Phylum, Class, Order, Family and Genus levels as well as inclusion of non-lesional samples. All differential analysis was subject to a parallel analysis using MaAsLin which was used to estimate the differential abundance of all taxa whilst controlling for potential confounding factors. Paired comparisons between lesional and non-lesional tissue were also incorporated. All implementation, interpretations, figures, and text was performed by myself.

3.2 Methods

3.2.1 Data acquisition and sampling

Raw and TMM normalised 16S rRNA sequencing reads were obtained from the Karolinska Institutet and Institut Curie as part of the MAARS consortium project. Briefly, consenting patients with mild-to-severe chronic AD and plaque type PSO and healthy volunteers were recruited from three university hospitals in Dusseldorf (HHU), London (KINGS) and Helsinki (UH). Diagnosis was made by a dermatologist using the Hanifin and Rajka criteria. Patients were excluded based upon antibiotic use and presence of autoimmune diseases. Swabs were taken from the skin surface using a ring and glass rod in 1.5 ml of PBS which was swabbed 10 times to the right and 10 times to the left. The 16S rRNA gene was amplifed from DNA extracted from surface swabs with PCR and sequenced on a Genome Sequencer FLX-Titanium instrument in Stockholm at the Karolinska institutet. After demultiplexing and removal of chimaeras, QIIME [95] was used to pick open reference OTUs at 99.3% identity. Taxonomy was assigned to OTUs with blast-2.2.22 [93] with the Greengenes reference database [92]. Three samples of poor quality were removed from the dataset. Raw OTU counts were normalised using the Trimmed Mean of M-values method (TMM) [102]. A detailed description of the patient recruitment and sampling performed by the MAARS consortium is described in (**Section 2.3.1**), and for 16S rRNA sequencing and microbiome processing refer to (**Section 2.3.2**).

All available samples in the MAARS dataset were considered in the analysis; however, a body site-matched cohort was also defined. This analysis considered a subset of healthy and disease samples which were more balanced with respect to body sites. AD and PSO tend to occur at different body sites, therefore, for analysis of the AD and healthy control (CTRL) microbiota, the matched cohort consisted of samples originating from the thigh and upper back which accounted for 98% of the available ADL samples and 52% of the available CTRL samples. The PSO matched cohort consisted of samples from the lower and upper back accounting for 85% of available PSOL samples and 54% of CTRL samples. An overview of the whole cohort and the matched AD and PSO cohorts are shown in **Table 3.1** and **Tables 3.3** and **3.4** respectively.

3.2.2 Species Diversity

The Chao1 richness estimator [104] and Shannon diversity was calculated within R package Vegan [162] with the functions estimateR() and diversity(). Calculation of community dissimilarities (β-diversity) was performed with Bray Curtis at all taxonomic levels. NMDS was performed within Vegan [162] on Bray Curtis dissimilarities with the function metaMDS(). For calculation of weighted and un-weighted UniFrac distances [107], samples were rarefied to an even sequencing depth of 3500 reads within the phyloseq package [163]. Aerobic or anerobic status for OTUs was characterised by Jens Schröder with reference to Bergey's Manual of Systematic Bacteriology [164].

3.2.3 Statistical analysis

Association of taxa with body site and sampling institution was performed with Kruskal-Wallis ranked sum tests. Associations with age were calculated using Spearman's correlation, and tests involving gender were performed with Wilcoxon ranked-sum tests. To determine if species diversity or species richness was associated with clinical group or global assessment score, Kruskal-Wallis tests were applied and post-hoc Dunn's tests were performed to identify significant pairs in the case of a null rejection. Unless otherwise stated, significant associations were considered as those with a Benjamini Hochberg adjusted p value < 0.05.

3.2.4 Differential abundance analysis

OTU counts were summed at the Phylum, Class, Order, Family and Genus levels where possible and relative abundances were calculated. OTUs not annotated to a specific taxonomic level were binned into an unassigned group. At each taxonomic level, a filtering step selected features that were non-zero in at least 10% of samples and with a minimum mean relative abundance of 0.001. Differential analysis between cohorts was performed with two approaches. In the first, a standard non-parametric Wilcoxon's ranked sum test was applied to the TMM normalised counts. In the second, linear models implemented in the MaAsLin package [5] were applied to taxa relative abundances. MaAsLin was used to estimate the disease effect whilst controlling for potential confounding factors using the formula: OTU $\sim$ clinical group + gender + anatomical location + sampling institution + age. OTU relative abundances were arcsine square root transformed [5]. Differentially

abundant taxa were identified as a consensus set, i.e., those significant with a Benjamini-Hochberg adjusted p value of < 0.05 by both MaAsLin and Wilcoxon's tests. Analysis between lesional and non-lesional pairs within disease groups was performed using paired Wilcoxon tests and with patient as a random covariate in MaAsLin. Unless otherwise stated, the -log(pvalue) reported within figures is the signed -log(BH pvalue) from the MaAsLin analysis. This corresponds to the sign of the linear regression coefficient multiplied by -log_{10}(pvalue) to retain the direction of association.

3.2.5 Classification

A supervised learning approach based upon Random Forest classification models was implemented to identify discriminative sets of taxa. The pipeline was used to train models for two comparisons (CTRL vs ADL, CTRL vs PSOL) with the aim of identifying the features that best discriminate disease groups. To identify the most stable predictors of disease, under-represented species were removed by filtering out those that were present in less than 15 samples in each comparison. Next, to select the most discriminant OTU features, Random Forest feature selection implemented in the R package Boruta [165] was used under a 10 cross fold validation framework with 10 randomised repeats. Selected features were ranked by Boruta according to the variable importance Z score. OTUs with a mean Z score greater than 0.2 were considered for further analysis. The fold change between healthy and disease patients was calculated to identify depletion or increased abundance in disease. After selection of features, the classification model was trained within the R package randomForest [166]. The performance of the model was evaluated using the R package ROCR [167], and the mean area under the curve (AUC) across all folds was reported. NMDS with Bray-Curtis distance using the selected feature sets with the function metaMDS from the R package Vegan [162] was used for visualisation of the predictive features. Selection frequency was calculated as the percentage of times a variable was selected by Boruta across all folds over all randomisations.

3.2.6 Co-occurrence network analysis

Co-occurrence networks were constructed on all OTUs present in more than 5 percent of samples resulting in a core microbiota of 569 OTUs. Compositionality-robust correlations were calculated with SparCC [124] on the raw OTU counts, which calculates a corrected

correlation coefficient designed specifically for assessing the correlation between taxa in microbiome studies. The statistical significance of correlations was evaluated against an empirical null distribution obtained with 100 bootstrap iterations and p values were corrected using the Benjamini Hochberg method ($p < 0.05$, SparCC > 0.2). Network visualisations were generated in Cytoscape [168].

3.3 Results

3.3.1 Study population

Healthy and diseased patients were recruited from three clinical institutions as previously described (**Section 3.2.1**). After quality control, 87 matched atopic dermatitis lesional (ADL) and non-lesional (ADNL) samples, 224 healthy control (CTRL) samples from 117 individuals, and 128 matched psoriasis lesional (PSOL) and non-lesional (PSONL) samples remained in the analysis (**Table 3.1**). Overall, PSO samples corresponds to a higher percentage of males (79%) and may reflect reports suggesting that severe disease is more prevalent in males than females [169].

Table 3.1: MAARS cohort study population

		ADL	ADNL	CTRL	PSONL	PSOL
Patients (n)		87	87	117	128	128
Samples (n)		87	87	224	128	128
Gender (n)	Female	39	39	146	27	27
	Male	48	48	78	101	101
Anatomical Location (n)	Buttocks	0	0	0	18	18
	Lower Back	2	3	107	87	99
	Thigh	45	45	104	1	1
	Upper Back	40	39	13	22	10
Institution (n)	HHU	36	36	66	49	49
	KINGS	14	14	89	43	43
	UH	37	37	69	36	36
Age	Mean	43.7	43.7	35.8	48.7	48.7
	SD	14.6	14.6	14.2	13.4	13.4

3.3.2 Characteristics of the skin microbiota

A total of 7532 unique OTUs were identified across five cohorts. The resultant OTU table was 98.5% sparse indicating that the vast majority of OTUs are rare (**Figure 3.1 A**). Only 297 OTUs were present in more than 10% of the dataset, and most samples were sparsely populated (**Figure 3.1 B**) with each sample on average consisting of 143 non-zero OTUs. Control samples had the greatest number of unique OTUs found across clinical group and approximately double the number of those found in ADL lesional samples (**Figure 3.1 C**).

No genus-level taxa were present in all samples, except *Staphylococcus* in ADNL (**Table 3.2**). Only four genera were present in 95% of samples on healthy skin. The most common genus was *Corynebacterium*, followed by *Staphylococcus*, *Acinetobacter* and *Burkholderia* representing a core component of the skin microbiota.

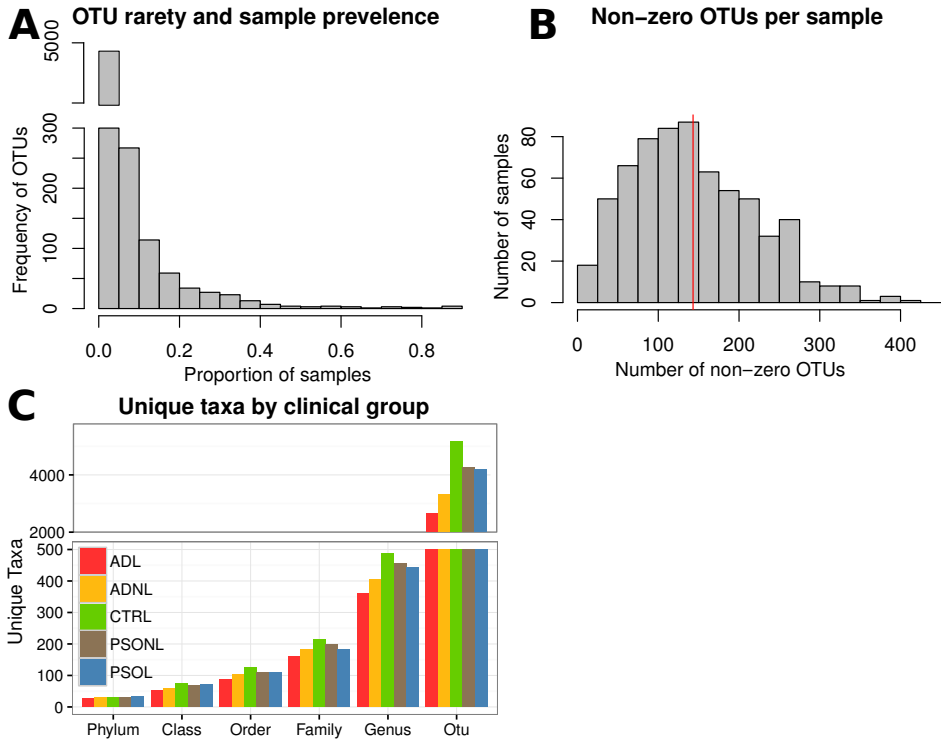

Figure 3.1: Microbiota summary statistics. (A) The frequency of OTUs present in a proportion of samples (with singletons removed). (B) Frequency of individual OTUs. The red line corresponds to the mean number of OTUs per sample of 143 (C) The number of unique taxa at each taxonomic level for each clinical group.

Table 3.2: Percentage occurrence of core genera across the MAARS cohort. (Genera present in at least 95% of control samples are shown)

	ADL	ADNL	CTRL	PSONL	PSOL
Staphylococcus	96.6	100.0	98.7	97.7	97.7
Corynebacterium	93.1	96.6	99.1	95.3	95.3
Acinetobacter	88.5	94.3	97.3	96.1	94.5
Burkholderia	83.9	93.1	97.3	97.7	95.3

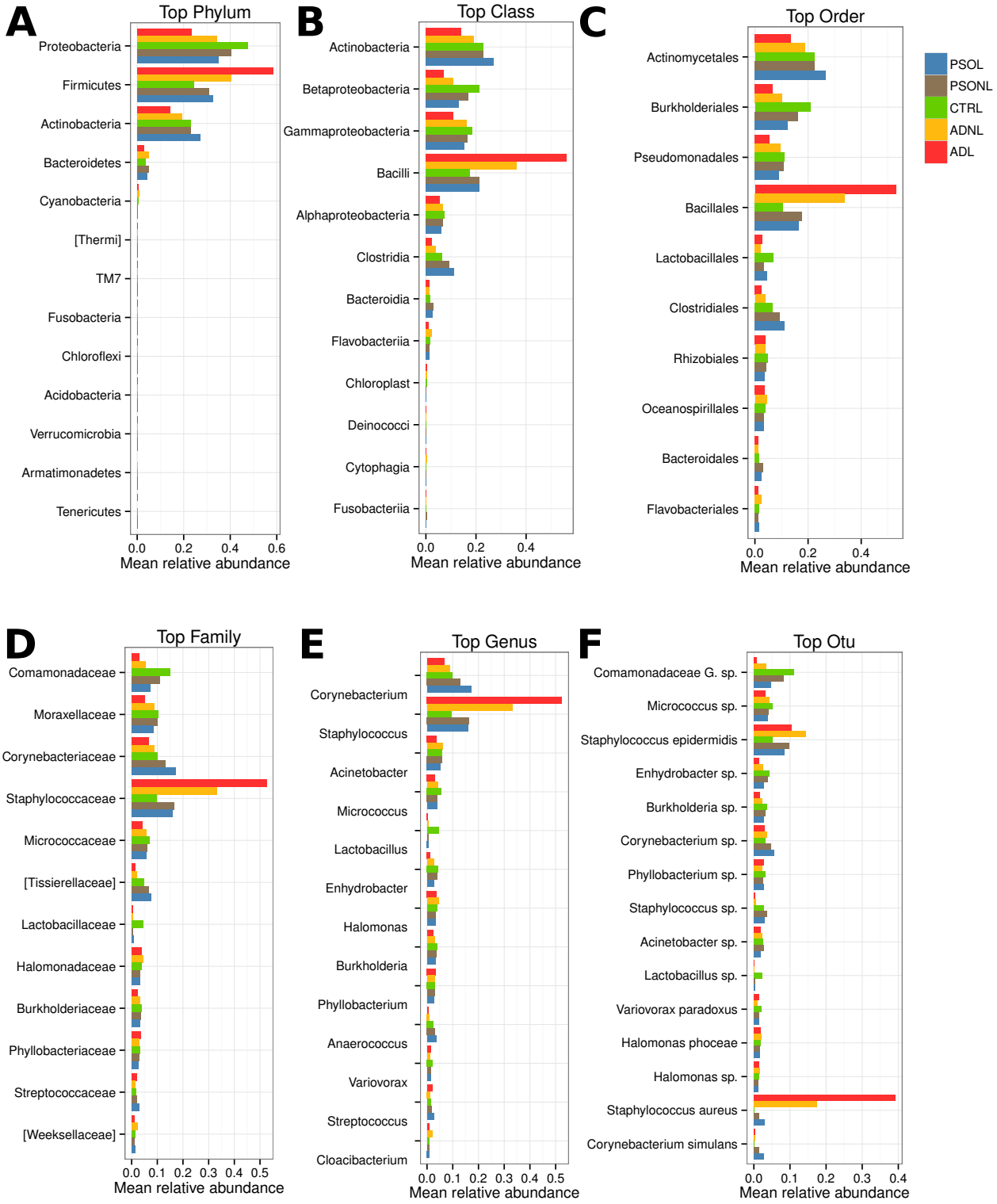

Figure 3.2: Abundant taxa at all phylogenetic levels across clinical groups.

To further define the core cutaneous microbiota, the most abundant taxa at each phylogenetic level and for each cohort were identified. Healthy skin was dominated by 4 main phyla (**Figure 3.2** A). *Proteobacteria* (47%), *Firmicutes* (25%), *Actinobacteria* (23%) and *Bacteroidetes* (4%) accounted for approximately 99% of the healthy skin microbiota which is in concordance with published analysis of the cutaneous microbiota [85, 84]. A major reduction in the relative abundance of *Proteobacteria* (23%) and a drastic increase in the abundance of *Firmicutes* (59%) was observed on atopic skin, whereas psoriasis was characteristic of moderate increases in *Firmicutes* (35%) and Actinobacteria (27%). A striking peak at each taxonomic level was observed in AD samples corresponding to *Bacilli*, *Bacillales*, *Staphylococcaceae*, *Staphylococcus* and *Staphylococcus aureus* at the Phylum, Class, Order, Family, Genus and OTU levels respectively (**Figure 3.2** A-F). Changes in PSO were much more subtle, and the most abundant taxa closely resembled healthy skin.

3.3.3 Community diversity in health and disease

3.3.3.1 Associations with α-diversity

The composition of a sample can be described in terms of α-diversity indices which represent the evenness and richness of a sample. The Shannon index and Chao1 species richness estimator [104] was calculated and Kruskal-Wallis was used to test for systematic differences in α-diversity between clinical groups. Post-hoc Dunn's tests revealed a significant difference in species richness between ADL and PSOL ($p = 0.0002$). No significant differences were identified between both diseases and control samples (**Figure 3.3 A**). Significant differences in Shannon diversity across clinical groups were identified ($p < 0.05$, **Figure 3.3 B**). Post-hoc tests revealed atopic samples were of reduced diversity compared to healthy microbiota (ADL: $p < 0.0001$, ADNL: $p < 0.0001$, **Figure 3.3 B**). As others have reported, no significant change in Shannon diversity was observed in psoriatic samples compared to healthy samples [86], however, diversity was significantly higher in PSO than in AD ($p < 0.0001$).

Studies of the microbiome have shown that in some conditions, disease severity can be linked to microbial diversity [110]. To test this hypothesis on the skin, Kruskal-Wallis tests within lesional disease revealed that species diversity was significantly associated with the physicians' global assessment score in both PSOL ($p = 0.007$, **Figure 3.3 C**) and ADL ($p = 0.02$, **Figure 3.3 D**). Post-hoc tests revealed that diversity was significantly reduced

in severe AD compared to mild disease (p = 0.008), and severe psoriatic lesions had lower diversity than both mild (p = 0.0024) and moderate disease (p = 0.008). These results indicate that severe disease exists in a heightened state of dysbiosis. Lastly, the abundance of anaerobes, aerobes and facultative anaerobes across clinical groups was evaluated. Both anaerobes and aerobes were of reduced abundance in AD (p <0.0001, **Figure 3.3 E-F**) as well as a vast expansion of facultative anerobes (p <0.0001, **Figure 3.3 G**). No significant associations were identified with O_2 tolerance status in PSO.

3.3.3.2 Associations with β-diversity

Next, compositional differences in the cutaneous microbiota (**Figure 3.4**) were evaluated between samples using measures of (β-diversity). Pairwise Bray Curtis dissimilarity was calculated between samples and ordination was performed with non-metric multidimensional scaling (NMDS) using the Vegan package [162]. At the OTU level, ADL and to a lesser extent, ADNL samples clustered with each other (**Figure 3.4 A**) indicating that the atopic microbiota is compositionally dissimilar from healthy and psoriatic communities. Involved and uninvolved psoriatic samples were indiscernible from healthy communities.

To determine if compositional differences on Psoriatic skin were could be better represented at a higher phylogenetic level, the analysis was repeated (ordination of Bray Curtis dissimilarities) at each taxonomic level, as well as the top 100 most abundant OTUs (**Figure 3.4 B-G**). In each case, ordination revealed a degree of clustering of amongst atopic samples which were compositionally dissimilar to the remaining clinical groups however, psoriatic communities were inseparable from healthy controls.

To determine if these findings were reproducible with different distance metrics, ordination of weighted and un-weighted UniFrac distances on samples rarefied to a depth of 3500 reads was performed. Unweighted UniFrac (**Figure 3.4 H**), which only takes into account phylogenetic relationships, showed little evidence of clustering amongst clinical groups, whereas weighted UniFrac (**Figure 3.4 I**) which takes into account both abundance and phylogeny, produced an ordination similar to those calculated with Bray Curtis dissimilarity. Therefore, these results indicate that the atopic microbiota is dissimilar to healthy and psoriatic communities, of which species abundance is an important factor, and that any compositional differences in psoriasis are subtle.

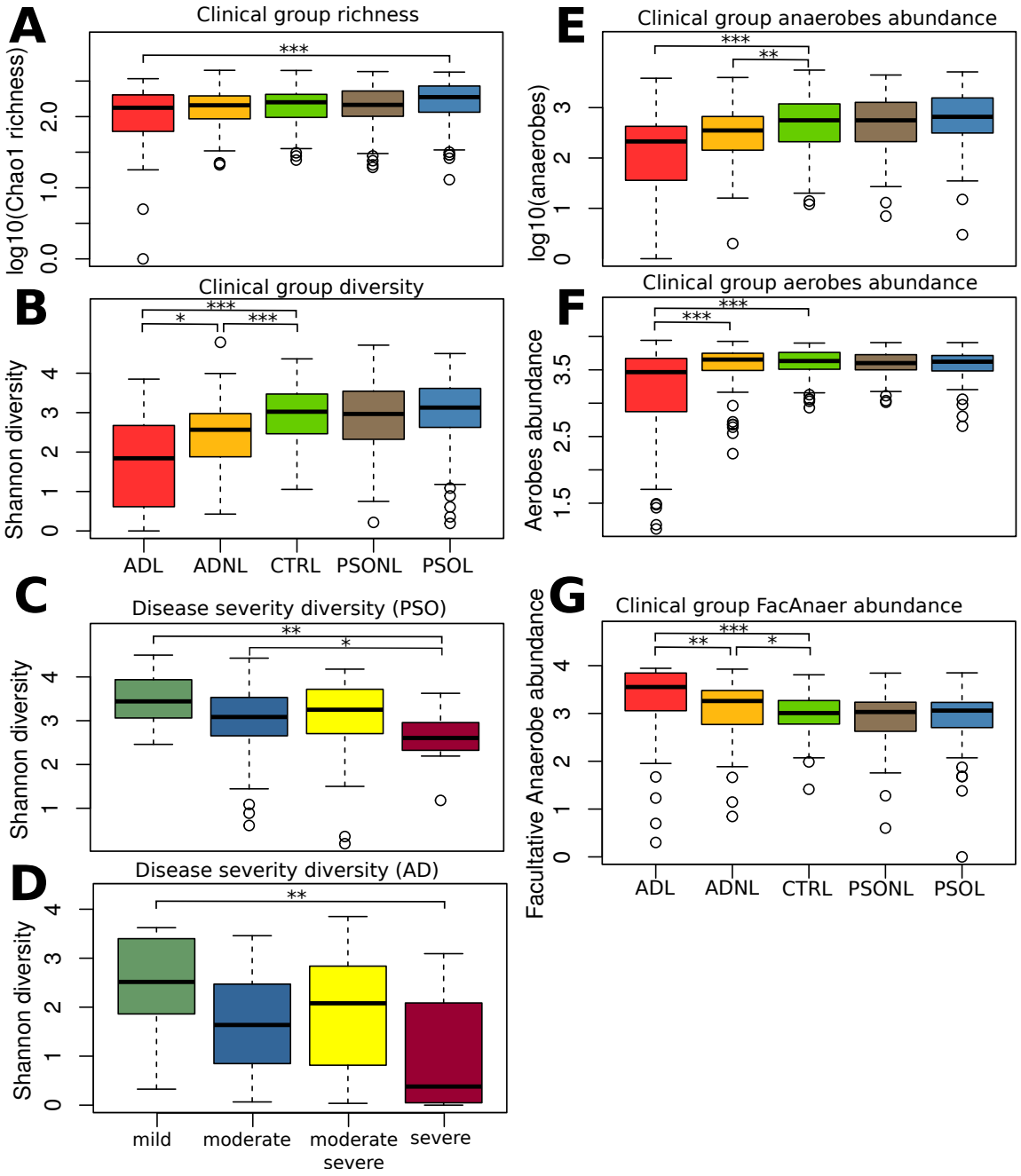

Figure 3.3: Clinical group association with species richness, diversity and microbiota O_2 tolerance. (A-B) Between-group comparison of α diversity. (A) Chao1 richness. (B) Shannon diversity. (C-D) Within-disease comparison of Shannon diversity across severity states (C) Shannon diversity stratified by global assessment score in PSOL samples. (D) Diversity stratified by severity in ADL. (E-G) Between-group comparison of abundance of specific O_2 tolerant species. (E) Association with anaerobes. (F) Association with Aerobes (G) Association with anaerobes. Stars represent Bonferroni corrected p value (Dunn's test, * p <0.05, ** p <0.005, *** p <0.0005)

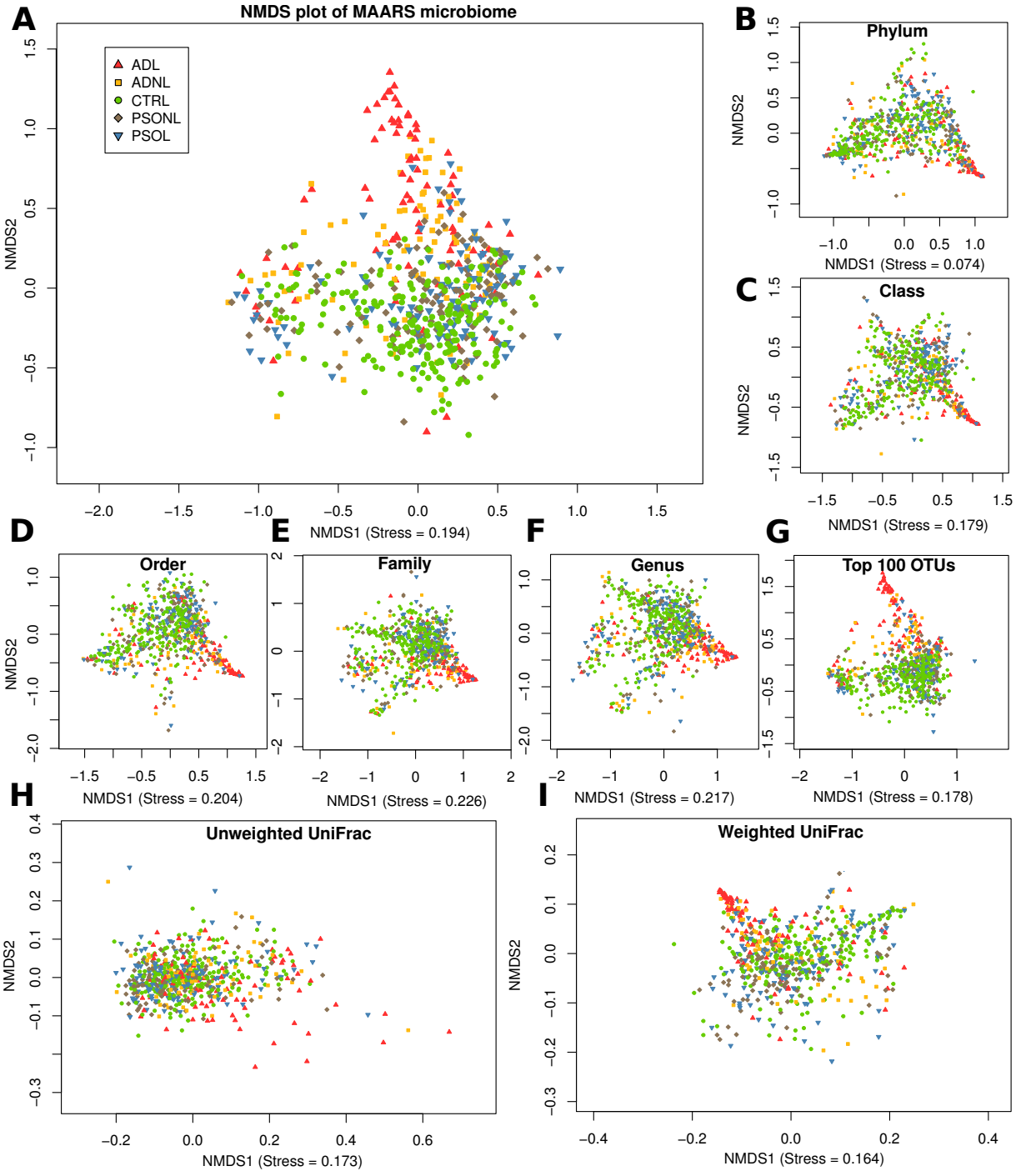

Figure 3.4: Community composition and β diversity. NMDS analysis of Bray Curtis dissimilarity of the MAARS cohort. Point colors and shape correspond to different clinical groups. (B-G) NMDS analysis of Bray Curtis dissimilarity at all phylogenetic levels and the top 100 most abundant OTUs. (H) NMDS of unweighted UniFrac distance on samples rarefied to 3500 reads. (I) Weighted UniFrac analysis.

3.3.4 Differential Abundance Analysis

A major objective of clinical microbiome studies is to identify species which are over-represented on diseased tissue. Before differential abundance analysis, a filtering step was first implemented to focus the analysis on highly abundant taxa. This step was performed as rare taxa are susceptible to under-sampling which results in spurious counts and could increase type I error. Further, filtering enables a reduction in the number of performed hypothesis tests, increasing the power to detect changes amongst abundant species. All OTUs were also aggregated at the Phylum, Class, Order, Family and Genus levels allowing for rarer taxa to contribute at higher levels of the taxonomic tree. Taxa at all levels were considered if they were present in 10% of samples, with a mean relative abundance of at least 0.001. Across all taxonomic levels, this filtering resulted in a total set of 285 taxa which were tested for differences in abundance. These included 8, 17, 30, 61, 65, and 104 taxa at the Phylum, Class, Order, Family, Genus and OTU levels respectively.

3.3.4.1 Non-clinical factor analysis

It is well known that microbial composition is associated with non-clinical factors such as age, gender [40] and body site [38]. To establish the relationship with non-clinical factors in the MAARS dataset prior to differential analysis, the changes in OTU abundance with respect to body site, gender, age and sampling institution were first evaluated at the OTU level using non-parametric statistics.

Of the 104 OTU level features evaluated, many were associated with non-clinical factors. Amongst the healthy controls, 28 OTUs were significantly associated with at least one factor ($p < 0.05$, **Appendix A.1**). Six OTUs were significantly associated with gender, all of which were of higher abundance in females, except for *Corynebacterium sp* which was of higher abundance in males. Of particular interest were OTUs mapping to the *Lactobacillus* genus for which 3 were significantly of greater abundance in females. Five OTUs were associated with body site including *Staphylococcus aureus* which was of higher abundance on the upper back, and *Finegoldia sp*, which was of reduced abundance on the upper back. Across all clinical groups, several significant associations were identified with sampling institutions indicating potential sampling bias, or that geographic location and or climate [117] may impact upon the abundance of some species.

In light of this observation, to identify a robust set of disease associated taxa, both

Wilcoxon's tests and linear models implemented in the MaAsLin package [5, 118] were applied (see Methods Section 3.2.4). As non-clinical factors were associated with taxa abundance, the factors age, gender, sampling institution and body site were included in the linear model to control for extraneous sources of variation. To minimise false positives, a consensus set was defined such that a taxon was only considered to be differentially abundant if it was significant (Benjamini Hochberg p <0.05) in both models. The models were then applied to identify differences in taxa abundances for ADL-CTRL, ADNL-CTRL, PSONL-CTRL and PSOL-CTRL.

3.3.4.2 Differential taxa at the phylum level

An evaluation of the phylum level taxa was performed first to identify higher-level variability between healthy and inflamed communities. At the phylum level, a gradient in the relative abundance of the three most abundant phyla across the whole dataset was observed (**Figure 3.5 A**) where some samples were dominated by *Proteobacteria*, and some by *Firmicutes*. A visual clustering of ADL samples towards the *Firmicutes*-high tail of the phylum distribution was observed suggesting that AD may be associated with increased abundance of *Firmicutes*. No distinct clustering of psoriatic or healthy samples was apparent. The expansion of *Firmicutes* and reduction of *Proteobacteria* in AD was also evident in the overall proportions of phyla stratified by clinical group (**Figure 3.5 B**). Wilcoxon ranked sum tests confirmed a significant reduction of *Proteobacteria* in both ADL (p = 1.12e-10) and ADNL (p = 4.62e-03) (**Figure 3.5 C**) along with significant increases in *Firmicutes* (ADL: p = 4.0e-14, ADNL: p = 5.67e-4, **Figure 3.5 D**). ADL was also characteristic of a reduction in *Actinobacteria* (p = 1.02e-05). Lesional psoriasis was associated with a moderate increase in abundance of *Firmicutes* (p = 5.58e-03) along with a reduction of *Proteobacteria* (p = 7.69e-05). No significant difference in *Firmicutes* or *Proteobacteria* was identified in non-lesional Psoriasis. A complete list of differentially abundant phyla is shown in **Appendix A**.

3.3.4.3 Differential taxa at the Class, Order, Family and Genus levels

Next, differences in taxa abundance were evaluated at lower taxonomic levels. At all taxonomic levels in atopic lesions, 106 taxa were found to be significantly different compared to healthy controls ($p < 0.05$). The most significant taxa were of increased abundance in AD including *Bacilli* at the class level (p = 8.83e-28), *Bacillales* at the order level (p = 3.22e-34), *Staphylococcaceae* at the family level (p = 3.22e-34), and *Staphylococcus* at the genus

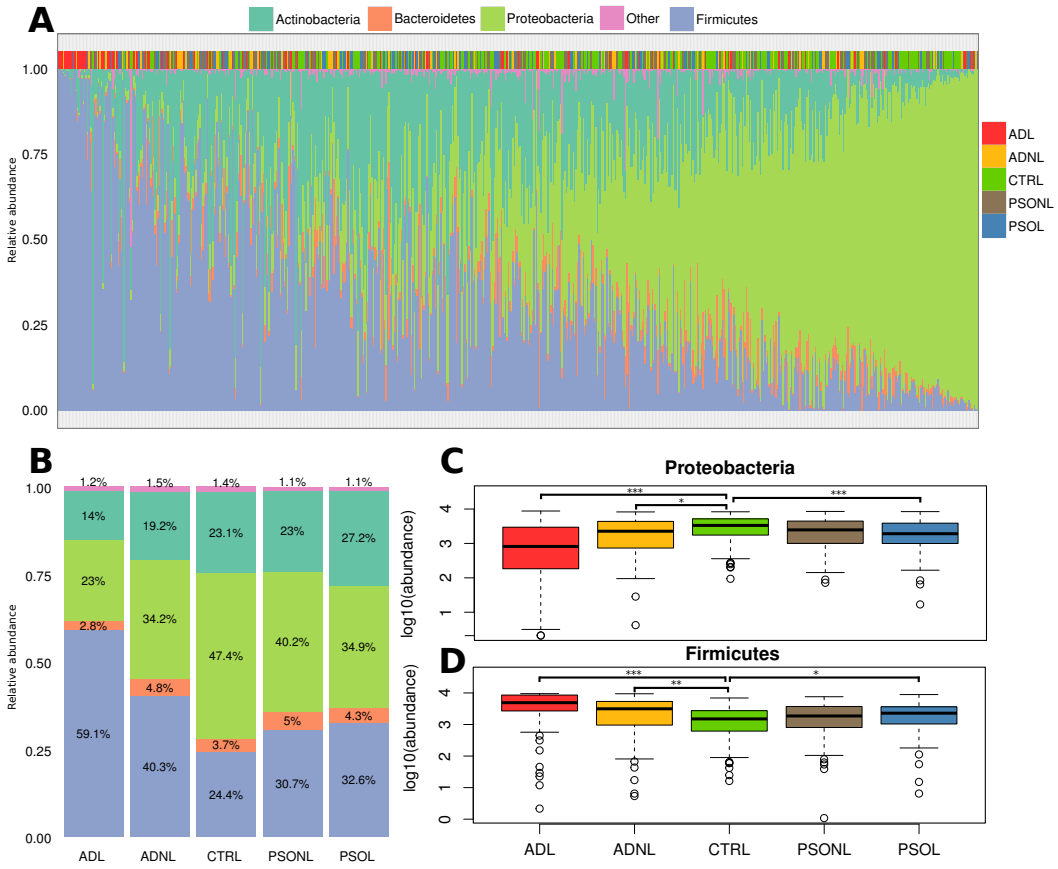

Figure 3.5: Relative abundances of the top 4 phyla. (A) Skyline plot showing the relative abundance of *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, *Firmicutes* and other phylum level taxa for each sample ordered by *Proteobacteria* abundance. The top of each column corresponds to clinical group. (B) Mean relative abundance of top 4 phyla. (C) Boxplot of *Proteobacteria* abundance across clinical groups. P values correspond to BH adjusted Wilcoxon ranked sum tests. (D) Differential abundance of *Firmicutes*. Stars represent BH corrected p value (Wilcoxon test, * p <0.05, ** p <0.005, *** p <0.0005)

level (p = 3.22e-34, **Figure 3.6 A-D)**. These results correspond to the well documented evidence of increased *Staphylococcus* abundance in atopic lesions [62]. The abundance of these taxa was also significantly higher on non-lesional skin compared to controls (**Figure 3.6 A-D)**. The vast majority of significant taxa were of reduced abundance on lesional AD skin (**Appendix A.2, A.3**). These included *Alpha* and *Betaproteobacteria* at the class level (p = 3.89e-10, 1.52e-08), *Burkholderiales* at the order level (p = 2.75e-08), *Burkholderiaceae* (p = 1.63e-06) and *Propionibacteriaceae* (p = 1.80e-05) at the family level, and *Burkholdiera* (p = 1.33e-06) and *Propionibacterium* (p = 2.19e-05) at the genus

level. These results suggest that a range of microbiota are of lower abundance on atopic skin and explain the dramatic loss of diversity. Many of the taxa found to be reduced on lesional skin were also of reduced abundance on non-lesional skin suggesting that the atopic microbiome is already altered in the absence of inflammation (**Appendix A.3**).

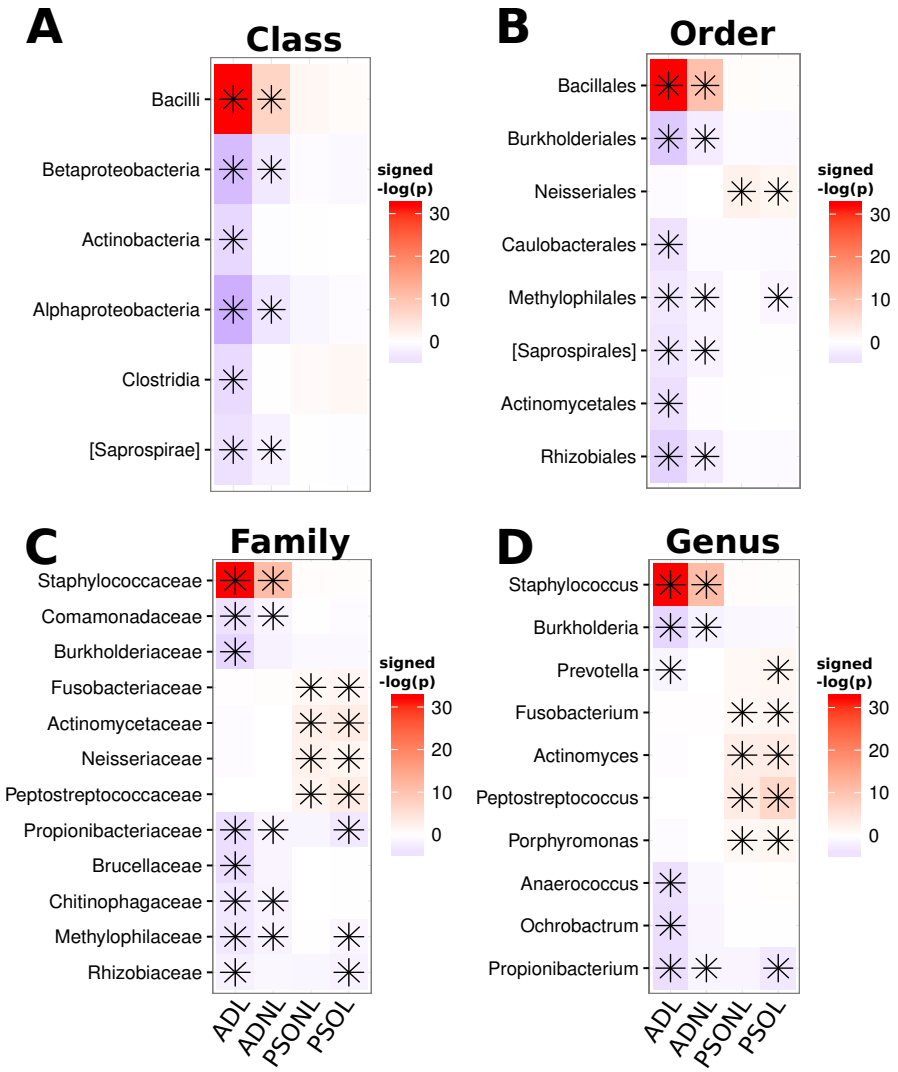

Figure 3.6: Differential abundance at the Class, Order, Family and Genus levels across clinical groups compared to control samples. Stars represent a significant association with both Wilcoxon and MaAsLin analysis. Tile color corresponds to the sign of the coefficient multiplied by -log10 of the adjusted p value. Positive values are of increased abundance in disease.

In comparison to AD, only thirty six taxa were found to be differentially abundant in PSOL, most of which were of greater abundance on diseased skin. At the Family level, the most significant included *Peptostreptococcaceae* (p = 5.12e-04) and *Actinomycetaceae* (p = 5.12e-04) which were of increased abundance in disease. At the genus level, *Peptostreptococcus* (p = 1.22e-07) and *Actinomyces* (p = 5.12e-4) were of increased abundance. As observed in AD, the most significant features were also differentially abundant on non-lesional skin (**Figure 3.6 C-D)** indicating that the uninvolved microbiota is already affected by changes to the microbial community. The most significant taxa of reduced abundance in PSOL included the family, *Propionibacteriaceae* (p = 5.12e-04) and the genus *Propionibacterium* (p = 6.27e-04). Very few taxa were differentially abundant in PSOL at the Class and Order levels (**Appendix A.4**) suggesting differences in microbiota are more specific. A complete list of significant taxa is shown in **Appendix A**.

3.3.4.4 Differential taxa at the OTU level

At the most specific OTU level, the microbial landscape in AD was dominated by *Staphylococcus aureus* which was over 150 fold greater in ADL compared to healthy (p = 1.72e-35, **Figure 3.7 A)**. As with other taxonomic levels, many OTUs were of reduced abundance in disease such as *Staphylococcus sp.*, *Burkholderia sp.*, *Corynebacterium sp.*, and *Propionibacterium acnes*. In non-lesional atopic skin, a significant and dramatic increase in the abundance of *S. aureus* compared to healthy was also observed (FC >50, p = 2.36e-16, **Figure 3.7 B**). Some of the species depleted on lesional tissue were also of reduced abundance on uninvolved skin including *Staphylococcus sp.*, and *Burkholderia sp.* Overall these results indicate that non-lesional communities in AD patients are already perturbed in absence of inflammation.

Several OTUs were of increased abundance in PSOL (**Figure 3.7 B**). The most significant was *Corynebacterium simulans* (p = 2.11e-09, FC >30). Other top psoriasis associated species included *Peptostreptococcus anaerobius*, *Neisseriaceae G. sp.*, *Streptococcus sp.*, two *Corynebacterium kroppenstedtii* OTUs, and *Peptostreptococcus sp.* In contrast to the atopic flora, only two species were found to be significantly depleted in disease by both methods including *Propionibacterium acnes* and *Burkholderia sp.* Many of the top psoriasis associated species were also significantly different in non-lesional tissue including *Corynebacterium simulans*, *Neisseriaceae G. sp.* and *Peptostreptococcus sp.* (**Figure 3.7 D**). Complete lists of differentially abundant OTUs are shown in **Appendix A**.

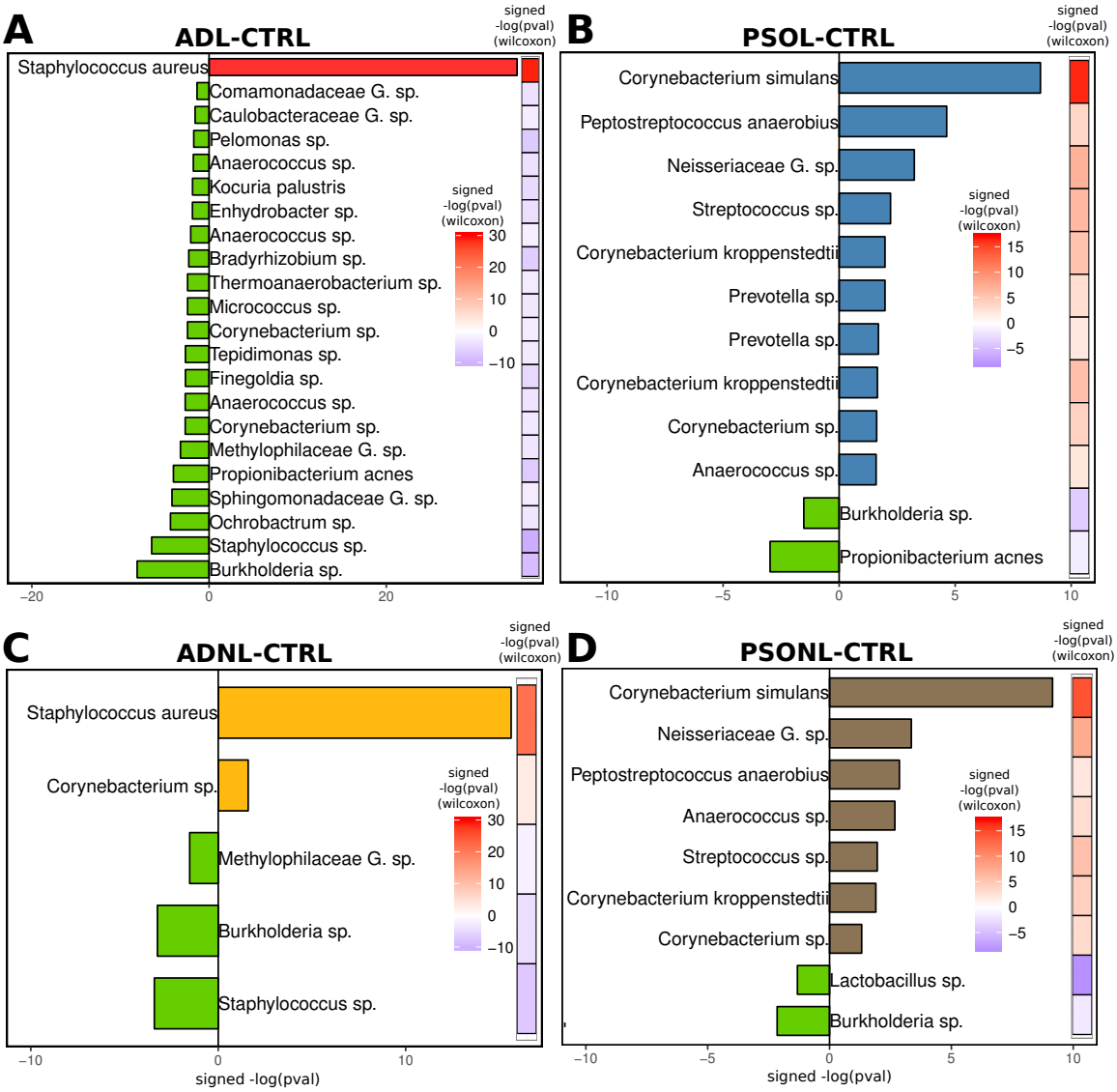

Figure 3.7: Differential abundance at the OTU level compared to control samples. OTUs shown were significant ($p < 0.05$) in both the Wilcoxon and MaAsLin analysis. Bars correspond to the sign of the coefficient multiplied by -log10 of the adjusted p value derived from the MaAsLin analysis. Heatmap tile color corresponds to the sign of the coefficient multiplied by -log10 of the adjusted p value from the Wilcoxon analysis. (A) Differential OTUs from the ADL-CTRL analysis. (B) PSOL-CTRL. (C) ADNL-CTRL. (D) PSONL-CTRL. Positive values are of increased abundance in disease.

3.3.4.5 Lactobacillus

Lactobacillus was one of the most significant species identified by Wilcoxon analysis. This taxa was found to be of reduced abundance in ADL (p = 3.17e-06), ADNL (p = 4.24e-04), PSONL (p = 8.85e-09) and PSOL (p = 9.31e-07), however, it was not identified as significant p <0.05 by MaAsLin after controlling for covariates (ADL p = 0.051, PSOL p = 0.07, **Figure 3.8 A**). *Lactobacillus* is clearly associated with gender and has a significantly higher abundance in female control samples (p = 2.4e-03, **Table A.1, Figure 3.8 B**). Whilst this species is confounded by gender, it is also heavily depleted across all cohorts (**Figure 3.8 A**). As studies have suggested a protective anti-inflammatory role for *Lactobacillus* in the gut [170], further study is warranted to determine if this species is indeed associated with homesostasis in the skin.

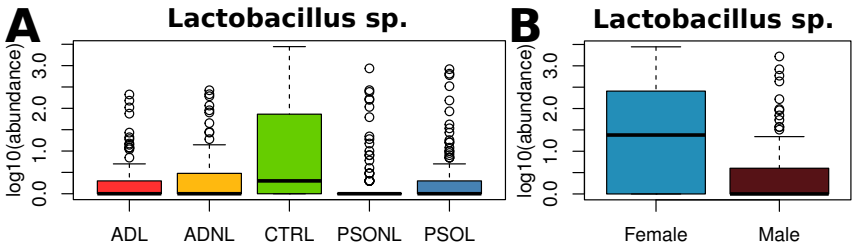

Figure 3.8: (A) Lactobacillus abundance across cohorts. (B) Lactobacillus abundance in CTRL samples stratified by gender.

3.3.4.6 Body site matched cohort

Disease associated taxa were identified by comparison to a healthy control baseline including all available samples; however, AD and PSO tend to occur on different body sites which is reflected in the samples collected for this study (**Table 3.1**). Further, as in some cases healthy patients contributed two or more samples across multiple body sites, it is important to determine the degree to which within-subject similarity influences the differential abundance analysis. The analysis was repeated considering only samples from disease specific matched body sites (see **Section 3.2.1**). AD samples were approximately balanced in the number of thigh and upper back samples and were compared to controls at the same sites (**Table 3.3**). The matched PSO analysis considered samples from the lower and upper back which accounted for 85% of PSOL samples (**Table 3.4**).

Table 3.3: AD body site matched cohort

		ADL	ADNL	CTRL
Patients(n)		85	84	113
Samples(n)		85	84	113
Gender (n)	Female	39	39	73
	Male	46	45	40
Anatomical Location (n)	Buttocks	-	-	-
	Lower Back	-	-	-
	Thigh	45	45	100
	Upper Back	40	39	13
Institution (n)	HHU	34	33	34
	KINGS	14	14	45
	UH	37	37	34

Table 3.4: PSO body site matched cohort

		PSOL	PSONL	CTRL
Patients(n)		109	109	115
Samples(n)		109	109	115
Gender (n)	Female	23	23	74
	Male	86	86	41
Anatomical Location (n)	Buttocks	-	-	-
	Lower Back	99	87	102
	Thigh	-	-	-
	Upper Back	10	22	13
Institution (n)	HHU	48	48	37
	KINGS	42	42	44
	UH	19	19	34

The majority of differentially abundant taxa were identified in both the matched and unmatched analysis except for a few differences. Considering ADL, 106 taxa were differentially abundance across all taxonomic levels in the unmatched analysis, and 105 in the matched analysis with an intersection of 96. Taxa which were not significant in the matched analysis (**Table A.6**) included an OTU assigned to *Anaerococcus* and the *Variovorax* genus. Regarding PSOL, the analysis identified 36 and 23 for the unmatched and matched analysis respectively, with an intersection of 20. Taxa which were not significant in the matched analysis (**Table A.7**) included *Corynebacterium kroppenstedtii* ($p = 0.063$), and *Prevotella sp* ($p = 0.059$). Furthermore, differences in *Firmicutes* ($p = 0.059$) and *Proteobacteria* ($p = 0.057$) were no longer significant ($p < 0.05$). Across both diseases, the most significant taxa were identified in both the matched and unmatched cohorts. Whilst

it's likely that some differences are related to body site and within-individual similarity, most species that lost significance were already towards the lower levels of significance in the unmatched cohort. This suggests that the differences are mostly due to a loss in power by removal of 30% of samples. To retain as much information as possible, and to benefit from having a single unified control cohort, the analysis was focused the entire dataset for further analysis. Complete lists of taxa which were not identified in the matched analysis are shown in **Appendix A**.

3.3.4.7 Differential taxa between involved and uninvolved cohorts

The cutaneous microbiota has been shown to be variable between individuals [39] therefore, differential abundance between involved and uninvolved tissue from the same patient can be performed to account for this. Only minor differences were observed between non-lesional and lesional AD. As others have found [64], the abundance of *Staphylococcus aureus* was significantly higher on atopic lesions compared to susceptible skin (p = 3.92e-07). The abundance of the *Proteobacteria*, *Actinobacteria* and *Bacteroidetes* phyla, as well as the *Finegoldia*, *Enhydrobacter* and *Micrococcus* (**Figure 3.9 A-F**) generas were all reduced on lesions. At the OTU level, the commensal *Staphylococcus epidermidis* was reduced on inflamed skin. No significant differences in microbial abundances were identified between PSONL and PSOL. This suggests that psoriasis patients are already subject to an altered microbial composition on uninvolved skin which does not drastically change during a flare.

The abundance patterns of the top disease associated microbes in AD and PSO across lesional and non-lesional disease was investigated further. Across paired AD samples, it can be observed that the mean relative abundance of *S. aureus* in ADNL was 0.17 and that 78% of pairs containing *S. aureus* increased in abundance on inflamed skin equalling a mean relative abundance of 0.39 in lesions (**Figure 3.9 G**). In psoriasis, *C. simulans* accounted for only 1.5% of the relative abundance in non-lesional tissue. Even though, 62% of pairs increased in *C. simulans* abundance across lesional status (**Figure 3.9 H**), the mean abundance on lesional skin accounted for only 3% of the psoriatic microbiota. It is therefore clear that the overall abundance profiles of the top pathogenic species are strikingly different and suggest that AD is dominated by one major species of which is not the case in psoriasis.

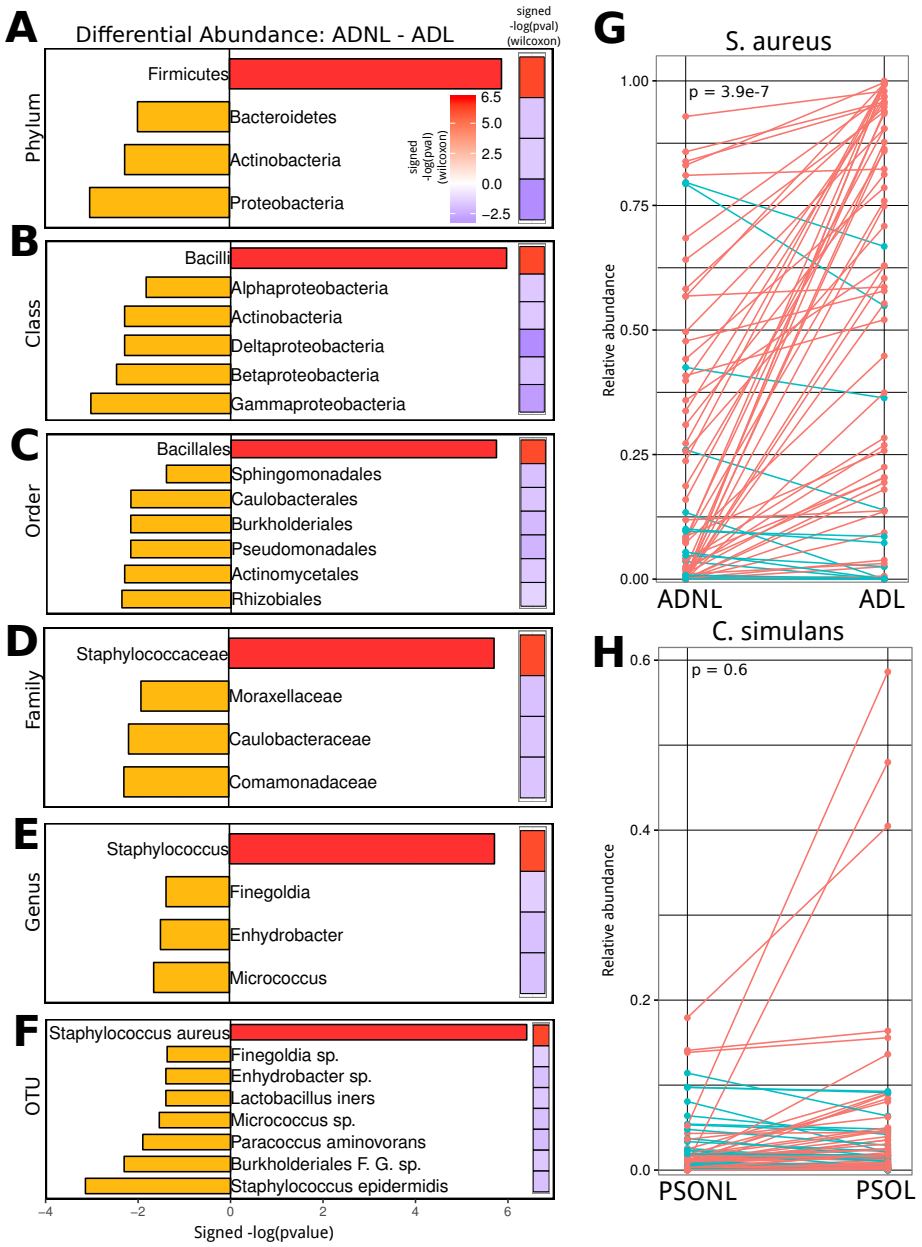

Figure 3.9: Differential abundance comparing uninvolved to involved skin in AD. OTUs shown were significant ($p < 0.05$) in both the Wilcoxon and MaAsLin analysis. Bars correspond to the sign of the coefficient multiplied by -log10 of the adjusted p value derived from the MaAsLin analysis. Heatmap tile color corresponds to the sign of the coefficient multiplied by -log10 of the adjusted p value from the Wilcoxon analysis. (A-F) OTUs of differential abundance across taxonomic levels. No significant associations were identified in PSO. (G) abundance of *S. aureus* across paired samples in AD. Red lines correspond to a patient which increased in relative abundance from uninvolved to involved. Blue lines correspond to those which decreased in abundance. (H) *Corynebacterium simulans* abundance in PSO. P values correspond to paired Wilcoxon ranked sum tests.

3.3.5 Classification and co-occurrence analysis

Next, to determine if the microbiota discriminates inflammatory skin pathologies, two classifiers were trained to explore the microbial taxa which differentiate diseased and healthy samples. A set of 25 microbes discriminating ADL and CTRL cohorts was identified with Boruta [165] to an AUC of 0.94 (class errors CTRL = 0.04, ADL = 0.26, **Figure 3.10 A**). The most discriminative microbial taxa were of the genus *Staphylococcus*, including *S. aureus* (Z = 14.0), *S. epidermidis* (Z = 5.8), *Staphylococcus sp.* (Z = -6.89) and *Burkholderia sp.* (Z = -7.4). Twenty six microbial taxa differentiated PSOL and CTRL with an AUC of 0.85 (class errors CTRL = 0.08, PSOL = 0.32, **Figure 3.10 B**). The top discriminating microbes were *C. simulans* (Z = 15.5), *Neisseriaceae g. sp.* (Z = 6.9), *C. kroppenstedtii* (Z = 5.5) as well as *Lactobacillus sp.* (Z = -8.4) and *Lactobacillus iners* (Z= -3.5).

To understand the interactions between communities of microbes under different disease states, network principles were applied to express co-occurrence relationships between disease-associated microbial taxa. Distinct differences between the community structures of microbes associated with ADL and PSOL were observed. For microbes associated with ADL, network inference with SparCC [124] resulted in 17 species with at least one significant interaction with another predictive feature (SparCC >0.2, p <0.05, **Figure 3.10 C**). *S. aureus* displayed a negative correlation with *Corynebacterium sp.*, *S. epidermidis*, *Tepidimonas sp.* and *Phyllobacterium sp.*

Of the 23 microbes identified as important for PSOL classification (**Figure 3.10 B**), 14 species showed significant correlation (SparCC >0.2, p <0.05 **Figure 3.10 D**). Fewer connections were observed in PSOL than in ADL and all correlations were positive. The most informative taxa, *C. simulans* and *C. kroppenstedtii* displayed positive correlations with *Streptococcus sp.*, *P. anaerobius* and, *Anaerococcus sp.*, *Neisseriaceae g. sp.*, *Neisseriaceae g. sp.* and *Rothia dentocariosa* respectively. Comparison across microbial interactions in PSOL suggests that, rather than a single species dominating the microbial landscape (as in ADL), multiple species are associated with this disease type. Lastly for visualisation purposes, ordination of OTUs identified by Boruta was performed with NMDS. In ADL, the top 25 classifying microbes highlights the role of *S. aureus* in differentiating ADL from CTRL samples (**Figure 3.10 E**). NMDS analysis of PSOL associated microbes demonstrated a separation boundary between lesional and healthy samples (**Figure 3.10 F**).

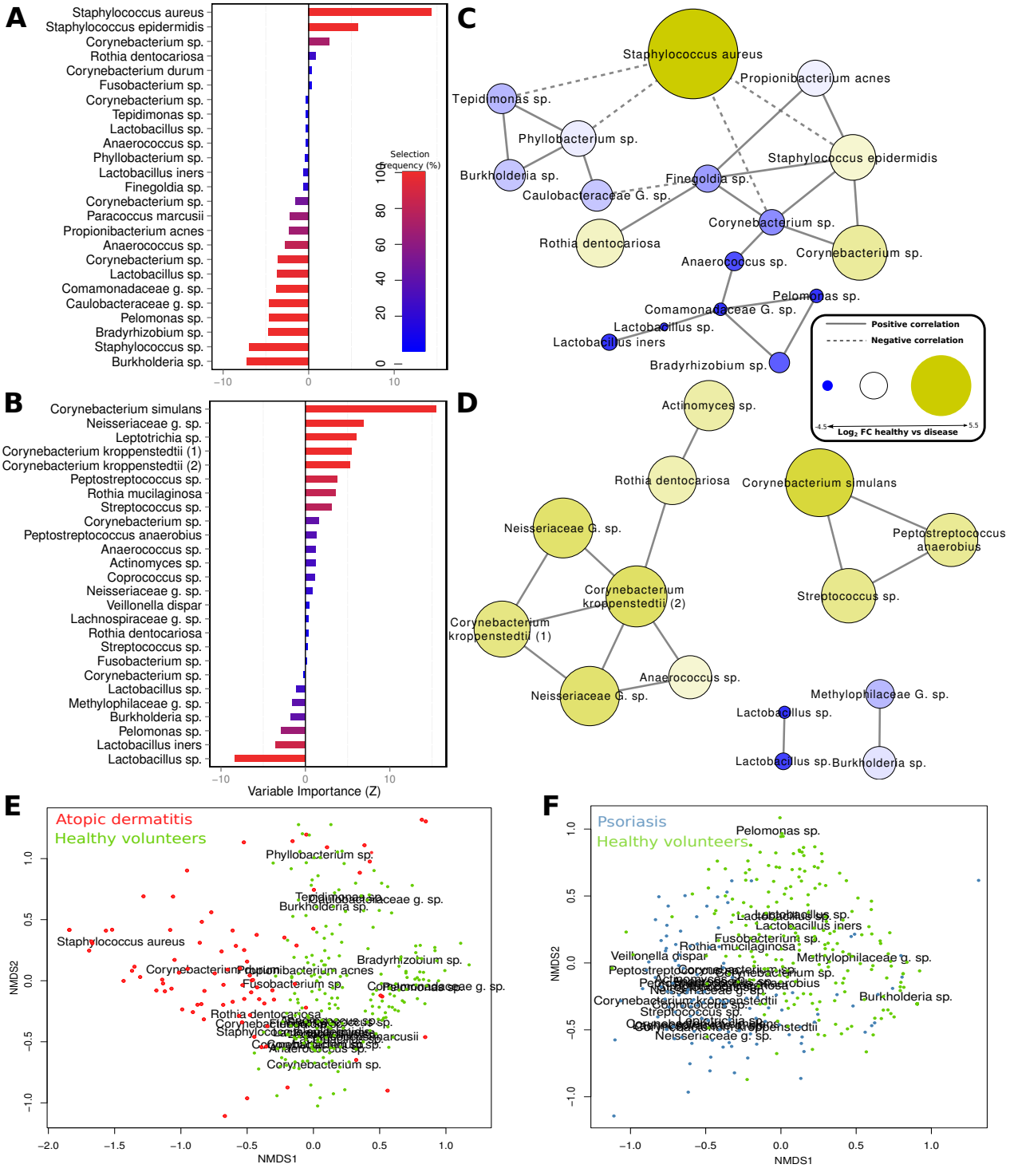

Figure 3.10: Classification models and co-occurrence network analysis. (A) Variable importance for the best set of discriminatory AD taxa identified with Boruta and classification analysis. Bars are coloured by selection frequency (Red = OTU selected in all folds, blue = OTU selected in 1 fold) (B) PSO taxa identified with Boruta. (C) Co-occurrence network of AD associated OTUs selected by Boruta infered with SparCC. A connection represents a correlation >0.2 and significant P <0.05 correlation. Solid and dashed lines respectively represent positive and negative correlations. The size and colour of each node is proportional to the log2 fold change between healthy and disease. (D) Correlation network of PSO associated OTUs. (E) NMDS analysis considering only microbes identified through feature selection for AD vs CTRL and (F) PSO vs CTRL.

3.4 Conclusions and Discussion

This chapter presents the largest analysis to date of the healthy and inflammatory skin microbiota. Overall, the skin harbours a complex and diverse ecosystem which is dominated by four main phyla consisting of *Proteobacteria*, *Actinobacteria*, *Firmicutes* and to a lesser extent *Bacteriodetes* confirming observations of previous studies in the cutaneous skin microbiome [85, 84, 38, 31]. As the main components of the microbiota, it is likely that these taxa exist in symbiosis with the human host to maintain a state of homeostatic equilibrium on the skin. Whilst the top 4 phyla were in concordance with previous studies, the overall proportions of these phyla conflict with results of a study which identified *Actinobacteria* as the most dominant phyla [84]. Instead, the results of this study found the most abundant taxa on healthy skin to be *Proteobacteria* which accounted for 47.4% of the healthy microbiota. Despite differences in overall initial proportions of the most abundant phyla, similar changes to their abundances was observed in disease. *Proteobacteria* was under-represented and the proportion of *Firmicutes* was over-represented in both diseases, although this was to a much greater extent in AD. Whilst changes in the abundance of key phyla were similar to those found in the Gao et al. study [84], the results presented here are the opposite to those found in psoriatic biopsies [86] indicating the need of further studies to define a clear microbial landscape at the site of skin inflammation.

Psoriatic skin is characteristic of extreme AMP expression [73], and atopic skin is subject to changes in pH and lower abundance of epidermal ceramides [44]. Therefore, given the extreme environmental differences between inflamed and healthy skin, it is surprising to find that samples did not cluster by clinical group at any taxonomic level. AD samples were to some degree distinguishable from healthy and PSO samples although no clear boundary existed, and the pattern more closely represented a gradient in community composition. Psoriatic samples were inseparable from CTRL samples so it is clear that differences in the atopic microbiome are more dramatic than those observed in PSO. Ordination of disease associated OTUs identified by supervised feature selection resulted in a clearer separation of clinical groups indicating that the skin microbiome does carry a signal of a diseased environment.

The major features the atopic microbiota corresponded to increased abundance of *Staphylococcaceae*, *Staphylococcus* and more specifically *Staphylococcus aureus*. *S. aureus* is a well characterised pathogen known to be associated with atopic inflammation [51, 62, 171].

Incredibly, the abundance of *S. aureus* was over 150 fold higher in lesional atopic skin compared to healthy and this species completely dominated the microbiota of some patients. Many more species of increased abundance were identified on psoriatic skin in this analysis than the previous largest study [85]. This could reflect differences in sampling techniques, body site locations, geographic location or intra-individual variability which is known to be a major factor influencing the cutaneous microbiota [39, 1]. As many of the differentially abundant species had low effect sizes, it is likely that a large cohort increased power to detect subtle differences in the cutaneous microbiota. Several taxa of increased abundance were found in PSO including the genera *Peptostreptococcus* and *Actinomyces*. The most significant OTU in PSO was *C. simulans* suggesting a potential pathogenic or opportunistic role of this species. Whilst *C. simulans* was consistently the most significant, it is unlikely that *C. simulans* can be considered a pathogen on a similar levels to *S. aureus*. Many AD samples were completely dominated by *S. aureus*, whereas the average relative abundance of *C. simulans* in PSO was approximately 3% of the total microbial composition. Other species including *Peptostreptococcus anaerobius* and *Neisseriaceae G. sp.* were also consistently over represented on psoriatic skin.

As well as *Propionibacterium acnes* which has been found to be under-represented in psoriasis [84], this species was also reduced on atopic skin. Furthermore, *Burkolderia sp* was also depleted in both diseases. These species may reflect beneficial species which are lost in disease, and coupled with the observation that species diversity was was reduced in severe disease, it could be that loss of protective species exacerbates inflammation. Further analysis could be performed to determine if these species interact with the host and express anti-inflammatory properties, or if they provide resistance to pathogen colonisation.

A widespread depletion of *Lactobacillus* in all cohorts to high significance compared to healthy via Wilcoxons tests was observed (ADL p = 3.17e-06, ADNL p = 4.24e-04, PSONL p = 8.85e-09 and PSOL p = 9.31e-07). In fact, *Lactobacillus sp.* was the most significant OTU of reduced abundance in PSO. In section 3.3.4, *Lactobacillus* was found to be significantly higher on female healthy skin compared to males. Consequently, the significance of *Lactobacillus* was heavily impacted when controlling for extraneous sources of variation within MaAsLin suggesting that the gender effect, and possibly other confounding factors are associated with this species abundance. Given that the MAARS dataset is unbalanced in terms of males and females, further analysis should be performed to identify if this

species is associated with disease as it has already been implicated as a beneficial microbe in the gut microbiome [170], and could play a protective role on skin.

One of the most surprising observations was that uninvolved skin shared many traits of inflamed disease skin. Only moderate differences between ADNL and ADL were detected, the most significant being the increase of *S. aureus* on lesional skin, and that 78% of patients displayed increased abundance in a lesional phase. It is therefore possible that *S. aureus* may be associated with, or trigger an inflammatory event [64]. As well as increased abundance of *S. aureus*, several species were of reduced abundance on lesional skin. Coupled with the observation of reduced species diversity in lesions and severe disease, these findings raise the possibility of probiotic treatments in AD to improve species diversity. Interestingly, no significant differences between uninvolved and inflamed skin in psoriatic patients were identified highlighting that the microbiota is already in a state of dysbiosis in the absence of inflammation. One possible reason could be that changes in the microbiota during an inflammatory phase are systemic and overwhelm the healthy resident community. As disease patients in the MAARS cohort were sampled during a lesional phase, systemic changes in the microbiome may have already occurred. Further work could investigate this by sampling diseased patients in a non-lesional phase to determine if their microbial composition more closely resembles healthy skin.

A progressive reduction of species diversity from healthy, through uninvolved, to lesional atopic skin was observed. It is likely that this reduction is mostly related to the extreme abundance of *S. aureus* which may overwhelm the resident commensal microbiota. Indeed, it has been shown previously that *S. aureus* correlated with species diversity [64]. With co-occurrence network analysis, it was clear in AD that *S. aureus* dominated the microbial landscape and negatively correlated with several skin commensals such as *S. epidermidis* and *Corynebacterium sp.* These results indicate that this dominant pathogen could be associated with the displacement of potentially regulatory or protective microbes. In contrast to AD, multiple species which possibly organise into communities is a more representative model of the psoriatic microbiota.

Taken together, this analysis presents a global picture of the skin microbiota across healthy skin and two models of cutaneous inflammation. The healthy microbial landscape was characterised and the components which were over-represented or depleted in disease were

defined. This exploratory analysis places AD as a disease characteristic of loss in microbial diversity which could be mediated by infection of *S. aureus*. On the other hand, microbial differences in PSO whilst present, were considerably less extreme. Little evidence was observed for dysbiosis, although this may be a factor in severe disease. Key potential species in PSO corresponded to *C. simulans*, *C. kroppenstedtii*, *Neisseriaceae G. sp.*, and *Peptostreptococcus anaerobius*. Overall, this chapter defines key species of importance and lays the foundation for study into host-microbe interactions.

Chapter 4

Transcriptomic profiles of skin inflammation

4.1 Introduction

Comparing levels of gene expression by differential analysis is a powerful tool for identifying core changes in the underlying transcriptome by finding differentially expressed genes (DEGs). DEGs provide a global view of the systematic differences between cohorts and can act as an entry point into understanding the biological processes which are perturbed in disease. Genes of changed expression may be drivers of disease and are potential targets for future study.

Several transcriptome analyses have been performed to compare the AD transcriptome to baseline healthy samples [172, 61, 173, 174]. The most striking differences have been identified in immune response, particularly relating to expression of Th2 cell products [61]. This is also supported by differences in chemokine expression [174], altered IL36, and upregulation of the Trem1 pathway signalling [173]. Other major differences have been identified within epidermal compartment [61, 173] indicating that barrier weakness is a major component of atopic inflammation. Researchers have also mined the transcriptiome to identify changes which may correspond to the 'dryness' phenotype which accompanies AD [171]. These include Olsson et al. [172] who identified differences in the expression of the epidermal water retention gene aquaporin 3 (AQP3) and Plager et al. [174] who suggested a role for lipid metabolism, indicating that PPARγ trended to be downregulated in AD. A recent meta analysis [58] combined the datasets from several independent

studies which increased statistical power to detect differences. Whilst a smaller number of genes were identified than the union of all individual analysis, the authors argued that the meta-analysis derived gene list is more biologically relevant and robust than any of the individual counterparts. They found that the atherosclerosis signalling pathway was significantly perturbed indicating that vascular inflammation is a component of AD. Whilst these studies highlight the major areas of interest, they are limited by very small sample sizes.

The psoriatic transcriptome has been analysed several times by microarray and RNA sequencing [175, 176, 177, 178, 179, 180, 181]. Most of these studies have identified increased expression of IFNG and interferon inducible products [175, 176, 178, 180], thus, implicating Th1 cells as a central component of psoriatic inflammation. Furthermore, transcriptome studies have provided solid evidence for the involvement of Th17 cells by identifying consistent up-regulation of IL17A and IL17 inducible transcripts [176, 175, 179, 180]. One of the most striking observations is a heightened level of innate immunity as measured by extreme levels of antimicrobial peptide expression [176, 177, 179, 180]. Another characteristic of psoriatic transcriptomes is dysregulation within the epidermal compartment [178, 177], particularly within the small proline rich protein (SPRR) and late cornified envelope (LCE) families which is likely to result in structural weaknesses. Defects in lipid metabolism pathways has also been identified by several studies [176, 177, 180] and may relate to increased permeability in the skin barrier, or to comorbidities such as metabolic syndrome which often accompanies psoriasis [182]. Two studies have linked psoriatic inflammation to Wnt signalling [181, 176]. Like AD, a meta-analysis combining 5 individual transcriptome studies identified a core robust transcriptome that consisted of a strong immune signature and enrichment for atherosclerosis signalling, and fatty acid metabolism [140].

Several analyses have focused directly on the differences in transcript expression between atopic and psoriatic skin. These comparisons have identified several transcriptomic components which differ between diseases in comparison to healthy samples. The overwhelming trend across these comparative studies indicate changes in the expression of T helper cell signatures placing AD and PSO at opposing ends of the Th1/Th2 axis [183]. In this early study, differences in chemokine expression were identified, emphasising that the CXC chemokine family were expressed in PSO which attract Th1 and neutrophils, and the CCL

chemokines were expressed in AD which attract Th2 cells. Interest into the role of Th17 cells in psoriasis increased [73] which challenged the hypothesis that psoriasis was primarily a Th1 associated disease. Guttman-yassky et al. [184] found that psoriasis had higher expression of the IL-23/Th17 pathway which was also corroborated by others [185, 186]. Psoriasis has been identified to have higher expression of certain antimicrobial peptides compared to AD [77, 185, 187]. Furthermore, studies have also shown that there are general expression differences in epidermal compartment, particularly within genes encoding the cornified envolope [188] suggesting that whilst barrier defects are present in both diseases, differences do exist. More recently, Guaranta et al. [185] compared a cohort of patients with coexisting disease, i.e., patients suffering with both AD and PSO which allowed comparisons to be performed within individuals increasing their power to detect biological differences. They found changes in the expression of metabolic transcripts, epidermal differentiation and further corroborated differences in T cell signatures showing heightened expression of Th17 and Th2 cytokines in PSO and AD respectively. The most recent study implicated IL36 as a psoriasis associated biomarker compared to other skin diseases [187] potentially implicating this cytokine in the immunopathogenesis of PSO.

These studies have provided a basis for the main transcriptomic differences between AD and PSO, however, they were underpowered with maximum sample sizes of 30. The analysis performed in this chapter compares healthy, non-lesional and lesional skin with a cohort of almost 3 times the size of previous largest comparative study which increases power to detect subtle differences. Using the MAARS cohort, the transcriptomic profiles of atopic dermatitis and psoriasis was refined and contrasted in the largest cohort to date.

Transcriptomics analysis in this chapter builds upon collaborative unpublished works by Marine Jeanmougin. The original differential expression analysis contrasted healthy to diseased samples and compared the AD-lesional and PSO-lesional associated DEGs to identify disease specific and common signatures. The work presented in this chapter also includes comparisons of healthy to non-lesional tissue and non-lesional to lesional tissue. The model was modified to account for the non-independence of multiple samples contributed from the same patient and a fold change criterion was incorporated ensuring that the analysis is comparable to other published inflammatory skin transcriptomics studies. All analysis, figures, interpretation and text was performed by myself.

4.2 Methods

4.2.1 Data acquisition, sampling and processing

Quality controlled and RMA normalised expression data was obtained from Institut Curie as part of the MAARS consortium project. Briefly, consenting patients with mild-to-severe chronic AD and plaque type PSO and healthy voulenteers were recruited from three university hospitals in Dusseldorf (HHU), London (KINGS) and Helsinki (UH). Diagnosis was made by a dermatologist subject to the Hanifin and Rajka criteria. Patients were excluded based upon antibiotic use and presence of autoimmune diseases. At the site of surface swabs intended for DNA collection microbiomics profiling, 6mm punch biopsies were taken under local anaesthesia and stored in RNAlater. RNA was extracted with the RNeasy Fibrous Tissue mini kit (Qiagen) and hybridised to Affymetrix Gene ST 2.1 arrays. The arrayQualityMetrics [156] method was applied to identify array failures. Data was subsequently normalised using the Robust Multi-array average (RMA) method implemented in the affy package [157]. A detailed description of patient recruitment sampling is described in **Section 2.3.1**, and for transcriptome profiling refer to **Section 2.3.3**.

4.2.2 Principal component analysis

Principal component analysis of the scaled and preprocessed expression data was performed to identify major directions of variability within the transcriptome using the R function *prcomp*. The variable loadings represent the contribution of each gene to the axis of PC1 and PC2 and were used as input to a pre-ranked GSEA [147] analysis of GO biological processes [142]. The top process satisfying a Bonferroni corrected p value > 0.05 in both positive and negative directions were used as a guide to annotate the major directions of variability.

4.2.3 Differential analysis

Differential analysis was performed using preprocessed and RMA normalised expression data. Contrasts of ADL-CTRL, ADL-ADNL, ADNL-CTRL, PSOL-CTRL, PSOL-PSONL, and PSONL-CTRL were performed to identify genes associated with clinical groups. Within the limma framework [137, 138], a linear model was fit to each gene to estimate the change in expression between clinical groups. Gender, and sampling institution were included as fixed effects, and patient was included as a random effect to account for non-independence

of samples taken from the same individual. Differentially expressed genes were defined as those with a Benjamini Hochberg adjusted P-value of < 0.05 and $\log_2$ fold change LFC of > 0.58 (approximately equal to an absolute fold change of > 1.5).

4.2.4 Functional analysis

Genes identified as statistically significant were assessed for enrichment of canonical pathways and upstream regulators using Ingenuity pathway analysis (IPA) [143]. Enrichment of GO biological processes was performed with Enrichr [189]. Up-regulated and down-regulated gene sets were input into IPA and Enrichr separately. Disease specific gene sets were identified by comparing the gene lists of ADL-CTRL and PSOL-CTRL. Genes commonly up-regulated and down-regulated in disease were determined as the intersection, and those specific to each disease were identified as the difference. Oppositionally differentially expressed genes were defined as differentially expressed in both ADL and PSOL compared to CTRL and which LFC > 0.58, in opposite directions.

4.3 Results

A quality controlled and normalised transcriptomics dataset consisting of 83 ADL, 81 ADNL, 213 CTRL, 121 PSONL, and 120 PSOL samples was received from the MAARS consortium as described in (**Table 4.1**).

PCA was performed to visualise sample scores in three dimensions which demonstrated broad transcriptional differences between non-lesional and lesional groups (**Figure 4.1**). The first and second principal components explained 11.7% and 4.95% of the variability in the transcriptome respectively. In line with other studies [177], samples were progressively distributed along a gradient defined by principal component 1 (PC1) suggesting this is an important component associated with clinical group. This analysis revealed that lesional groups of both diseases were distinct from uninvolved groups, however, non-lesional and healthy samples were not linearly separable.

It is likely that PC1 represents high-level core transcriptomic differences between healthy and inflamed skin. To further investigate this, GSEA [147] was performed using the principal component loadings with GO biological process gene sets. As expected, the most significant gene set enriched in the positive direction of PC1 was ‘immune response’ (p

Table 4.1: MAARS Transcriptome study population

		ADL	ADNL	CTRL	PSONL	PSOL
Patients (n)		83	81	113	121	120
Samples (n)		83	81	213	121	120
Gender (n)	Female	36	37	135	25	26
	Male	47	44	78	96	94
Anatomical Location (n)	Buttocks	0	0	0	18	19
	Lower Back	2	3	99	80	90
	Thigh	44	42	100	1	1
	Upper Back	37	36	14	22	10
Institution (n)	HHU	34	32	58	45	44
	KINGS	13	14	86	41	41
	UH	36	35	69	35	35
Age	Mean	43.5	44.3	35.0	48.6	48.8
	SD	14.4	14.7	13.3	13.4	13.5

= 0.001) suggesting that the major transcriptomic gradient associated with inflamed skin relates to perturbations within the immune system. Cell substrate adhesion was enriched in the negative direction (p = 0.006) of PC1 towards healthy samples representing a state of homeostasis.

4.3.1 Differential gene expression analysis

A series of differential analyses were performed to identify genes which were dysregulated between inflamed, uninvolved and healthy skin. For both diseases, lesional disease was compared to healthy, as well as lesional to non-lesional and non-lesional to healthy using limma [137]. Differentially expressed genes were defined as those which were significant ($p < 0.05$) with a log fold change (LFC) > 0.58 (equalling a fold change of approximately 1.5). The numbers of differentially expressed genes is shown in **Table 4.2** and the top 10 most significant genes for each contrast is displayed in **Table 4.3**. After obtaining a global view, transcriptomic signatures were further analysed and dissected to identify gene sets preferentially expressed in AD or PSO in **Section 4.4**.

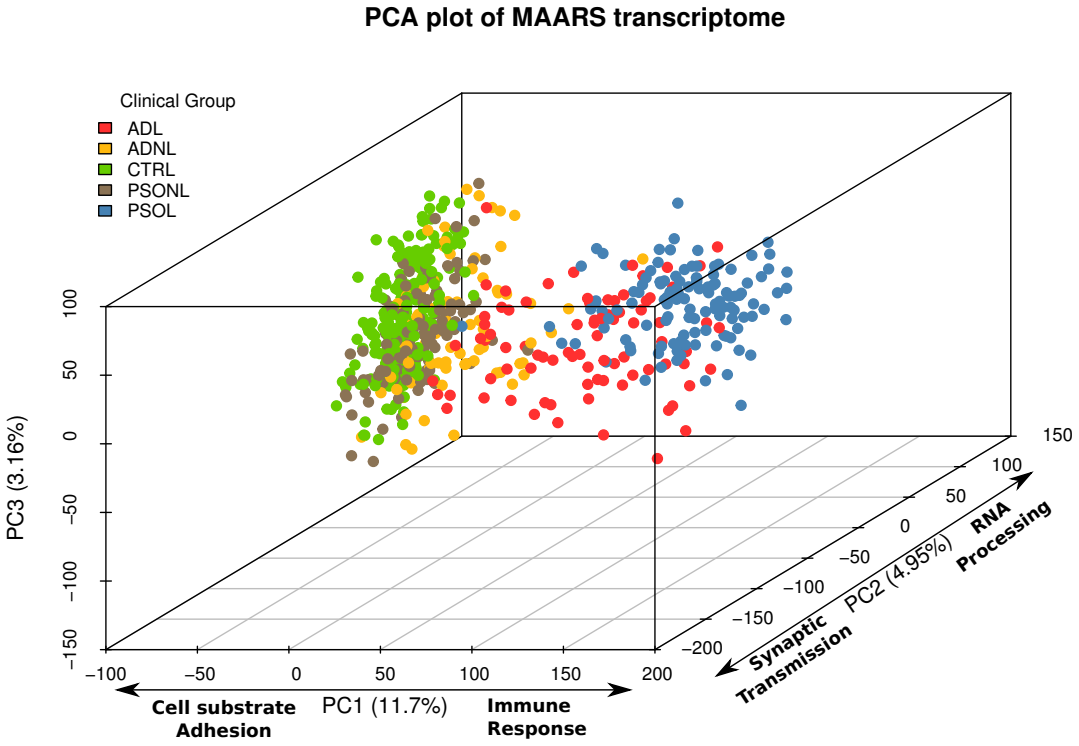

Figure 4.1: Principal component analysis of the MAARS transcriptome cohort. Coloured points correspond to different clinical groups. Terms along the arrows of PC1 and PC2 correspond to the top GSEA term identified using the ranked variable loadings.

Table 4.2: Differentially expressed genes (LFC > 0.58, p > 0.05)

Contrast	DEG	Up	Down
ADL-CTRL	1260	797	463
ADL-ADNL	500	353	147
ADNL-CTRL	86	55	31
PSOL-CTRL	2516	1326	1190
PSOL-PSONL	2217	1164	1053
PSONL-CTRL	26	23	3

Contrast	logFC	adj.P.Val	Symbol
ADL-CTRL	1.5E+00	1.6E-121	ARNTL2
ADL-CTRL	3.6E+00	5.7E-119	LCE3D
ADL-CTRL	2.9E+00	1.4E-118	AKR1B10
ADL-CTRL	3.7E+00	1.2E-116	LCE3E
ADL-CTRL	3.0E+00	1.2E-116	C10orf99
ADL-CTRL	3.3E+00	5.7E-116	LCE3A
ADL-CTRL	7.5E-01	7.3E-115	PPP4R1
ADL-CTRL	-2.3E+00	1.1E-113	BTC
ADL-CTRL	7.2E-01	2.3E-113	PGM2
ADL-CTRL	4.4E+00	4.4E-113	S100A9
ADL-ADNL	2.8E+00	5.8E-68	S100A12
ADL-ADNL	2.5E+00	5.8E-68	LCE3A
ADL-ADNL	2.1E+00	3.7E-66	AKR1B10
ADL-ADNL	2.9E+00	4.0E-61	S100A7A
ADL-ADNL	1.5E+00	1.0E-59	PRSS27
ADL-ADNL	9.5E-01	1.6E-58	LOC100507420
ADL-ADNL	2.6E+00	1.5E-57	KRT16
ADL-ADNL	9.6E-01	4.6E-56	CDH3
ADL-ADNL	9.6E-01	2.6E-54	GALNT6
ADL-ADNL	1.4E+00	3.6E-53	TMPRSS4
ADNL-CTRL	2.5E+00	7.1E-58	SPRR2G
ADNL-CTRL	2.0E+00	3.4E-50	LCE3D
ADNL-CTRL	2.4E+00	2.7E-48	S100A7
ADNL-CTRL	1.5E+00	5.2E-45	CCL13
ADNL-CTRL	1.1E+00	5.0E-38	TMC5
ADNL-CTRL	-1.8E+00	5.0E-38	WIF1
ADNL-CTRL	1.2E+00	2.0E-36	CCL18
ADNL-CTRL	1.4E+00	1.4E-34	C10orf99
ADNL-CTRL	-8.2E-01	1.4E-33	CHRM4
ADNL-CTRL	6.4E-01	5.0E-32	ARNTL2
PSOL-CTRL	3.7E+00	1.3E-286	KYNU
PSOL-CTRL	5.1E+00	2.0E-266	TMPRSS11D
PSOL-CTRL	6.4E+00	1.5E-237	S100A12
PSOL-CTRL	4.1E+00	8.3E-235	IL36G
PSOL-CTRL	5.7E+00	6.2E-229	IL36A
PSOL-CTRL	3.3E+00	3.8E-228	PLA2G4D
PSOL-CTRL	6.5E+00	5.2E-227	TCN1
PSOL-CTRL	6.6E+00	5.6E-223	S100A7A
PSOL-CTRL	2.6E+00	1.2E-216	KLK13
PSOL-CTRL	3.2E+00	1.4E-212	PRSS27
PSOL-PSONL	3.5E+00	1.6E-280	KYNU
PSOL-PSONL	5.0E+00	2.2E-268	TMPRSS11D
PSOL-PSONL	6.1E+00	1.3E-233	S100A12
PSOL-PSONL	5.7E+00	7.3E-233	IL36A
PSOL-PSONL	3.3E+00	4.8E-231	PLA2G4D
PSOL-PSONL	3.8E+00	9.7E-228	IL36G
PSOL-PSONL	6.1E+00	1.3E-222	TCN1
PSOL-PSONL	2.6E+00	6.4E-222	KLK13
PSOL-PSONL	6.2E+00	5.2E-216	S100A7A
PSOL-PSONL	1.9E+00	4.1E-215	VNN3
PSONL-CTRL	1.6E+00	1.1E-29	SPRR2G
PSONL-CTRL	1.6E+00	8.6E-27	S100A7
PSONL-CTRL	8.0E-01	1.7E-23	IFI27
PSONL-CTRL	-1.0E+00	1.5E-15	WIF1
PSONL-CTRL	5.9E-01	2.1E-14	TMC5
PSONL-CTRL	1.2E+00	1.2E-09	CXCL10
PSONL-CTRL	6.5E-01	1.6E-08	C10orf99
PSONL-CTRL	9.6E-01	2.3E-08	S100A9
PSONL-CTRL	9.3E-01	4.1E-08	S100A8
PSONL-CTRL	6.9E-01	5.8E-08	CHI3L2

Table 4.3: Top 10 differentially expressed genes for each contrast

4.3.2 Comparison of uninvolved skin between clinical groups

4.3.2.1 DEGs and pathways in non-lesional atopic skin

First, non-lesional skin was compared with healthy volunteers to identify genes which may convey disease susceptibility in absence of inflammation associated transcriptomic disorder. Considering non-lesional skin, 86 genes were differentially expressed (55 up, 31 down, **Table 4.2)** which were significantly dysregulated between ADNL and CTRL (**Figure 4.2 A**). Unlike [61], no significant downregulation of terminal differention genes FLG, IVL and LOR was found. Instead, other members of the epidermal differentiation complex were upregulated in non-lesional AD. These including members of the SPRR family, SPRR2B, SPRR2G and members of the LCE family, LCE3A, LCE3D and LCE3E. Antimicrobial peptides were also of increased expression in ADNL such as DEFB4A, S100A7, S100A7A and S100A9.

As well as differences in epidermal barrier gene expression, several genes associated with immune response were upregulated including AhR receptor translocator like 2 (ARNTL2), BIRC3, interferon alpha inducible protein 27 (IFI27), the IL1 family member IL36G, and genes associated with chemotaxis including CCL13, CCL18, CXCL9 and CXCL10. Ingenuity pathway enrichment showed that the top pathways were 'Role of IL17A in Psoriasis' which contained the S100A antimicrobial peptides and 'Granulocyte Adhesion and Diapedesis' consisting of CXC and CC family cytokines as well as MMP3 (**Figure 4.2 C**).

The top down regulated gene in ADNL was WNT inhibitory factor 1 (WIF1) indicating that wnt signalling may be perturbed in non-lesional skin. Several of the downregulated genes corresponded to lipid and fatty acid metabolism such as ACOT1, FABP7, AADACL3, AWAT2, and DGAT2L6 which suggests that deficiencies in lipid biosynthesis are pre-existing in non-inflamed skin.

4.3.2.2 DEGs and pathways in non-lesional psoriatic skin

Transcriptional activity within non-lesional psoriatic tissue was perturbed to a smaller extent and more closely resembled healthy skin with only 26 differentially expressed genes (23 up, 3 down, **Table 4.2, Figure 4.2 B**). In a similar manner to that observed in AD, the top genes included those associated with the epidermal differentiation complex including SPRR2G, SPRR2B, the antimicrobial peptides S100A7, S100A9 and S100A8 as well as

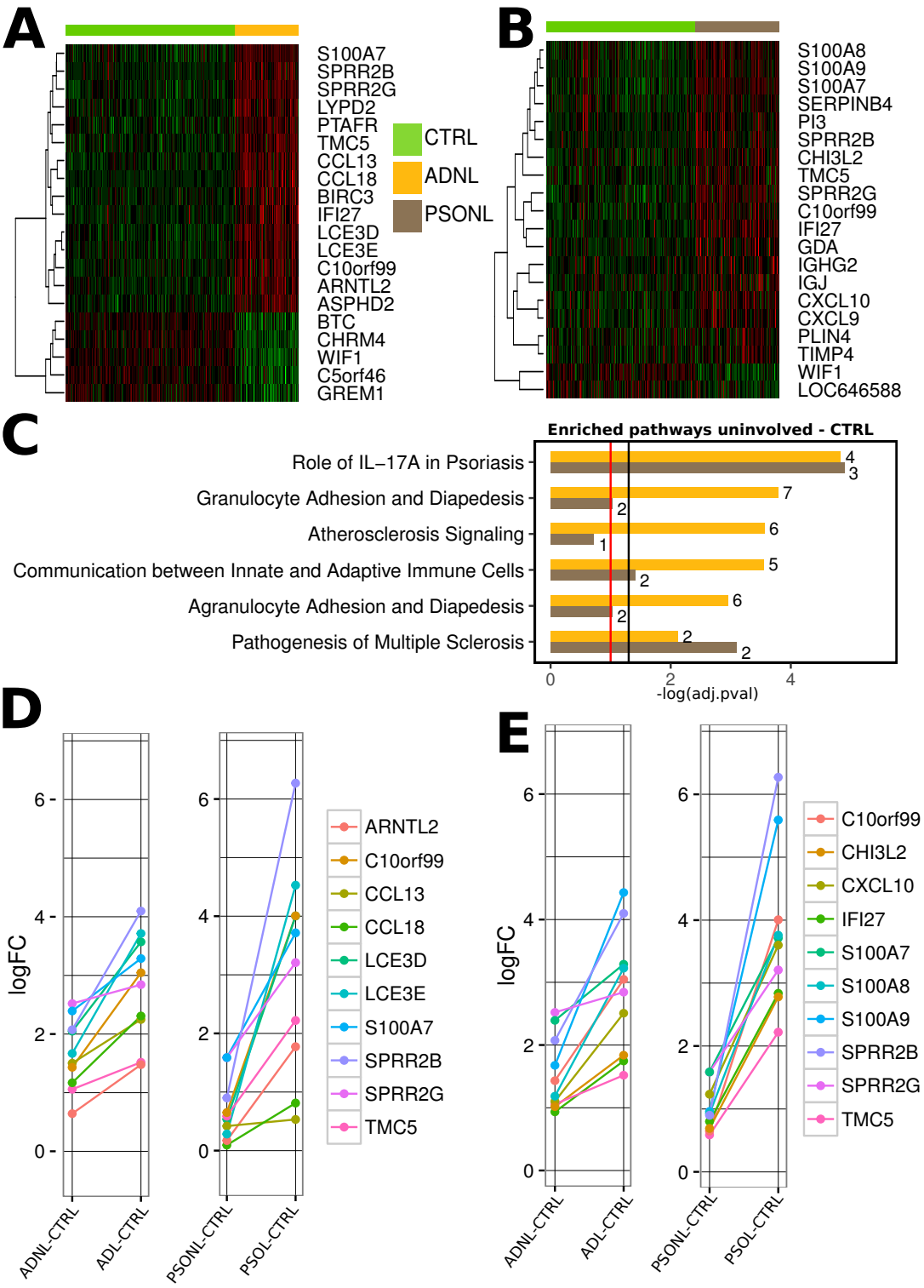

Figure 4.2: Genes differentially expressed in uninvolved skin compared to healthy. (A) - heatmap of top 20 differentially expressed genes between ADNL (orange) and CTRL (green). (B) DEGs between PSONL (brown) and CTRL (green). (C) Fold change comparison across uninvolved and lesional states. Top 10 genes from ADNL-CTRL. (D) Top 10 genes from PSONL-CTRL. (E) Enriched pathways amongst DEGs in both ADNL-CTRL and PSONL-CTRL contrasts. Pathways shown are significant (BH p value < 0.1) in at least one contrast.

immune genes including IFI27, CXCL9 and CXCL10.

A significant enrichment for the 'Role of IL-17A in psoriasis' was identified which, like AD, contained the S100A family antimicrobial peptides (**Figure 4.2 C**). Only 3 genes were downregulated in PSONL including WNT inhibitory factor 1 (WIF1). WIF1 is known to be strongly downregulated in lesional skin [190], however, this gene has not been described in non-lesional psoriatic [177] or atopic skin. Many of the genes dysregulated in PSONL were also significantly differentially expressed in ADNL, suggesting that skin susceptible to inflammation share a proportion of the same characteristics.

Comparison of the fold changes of significant genes shows that transcriptomic dysregulation within ADNL is 'more-extreme' than in PSONL (**Figure 4.2 D-E**). The fold changes amongst the top 10 DEGs compared to healthy were found to be higher in uninvolved atopic skin suggesting that non-lesional AD skin exists in a heightened state of disorder compared to non-lesional psoriatic skin. In a lesional state, this pattern is reversed where the same genes in PSOL skin were of greater expression.

Overall, analysis of non-lesional skin showed that there is a clear overlap in transcriptional changes observed in both diseases, as well as several key differences. Dysregulation in components of the EDC was observed in both diseases by upregulation of genes within the SPRR and S100A families. Further, common processes involving genes associated with chemotaxis such as CXCL9 and CXCL10 were upregulated indicating that uninvolved skin has heightened immune activation in the absence of inflammation. Further transcriptional dysregulation was observed in ADNL including the up-regulation of CC family cytokines, increased expression of LCE genes, and downregulation of genes involved in fatty acid metabolism.

4.3.3 Genes and pathways upregulated in lesional skin

4.3.3.1 Genes and pathways upregulated in lesional atopic skin

Next, differential expression was evaluated in lesional samples. A total of 1260 (797 up, 463 down, **Table 4.2**) DEGs between ADL and CTRL samples were identified (**Figure 4.3 A, B**). Within the top DEGs, several known genes previously identified in published transcriptomes were identified. These included genes corresponding to the epidermal barrier

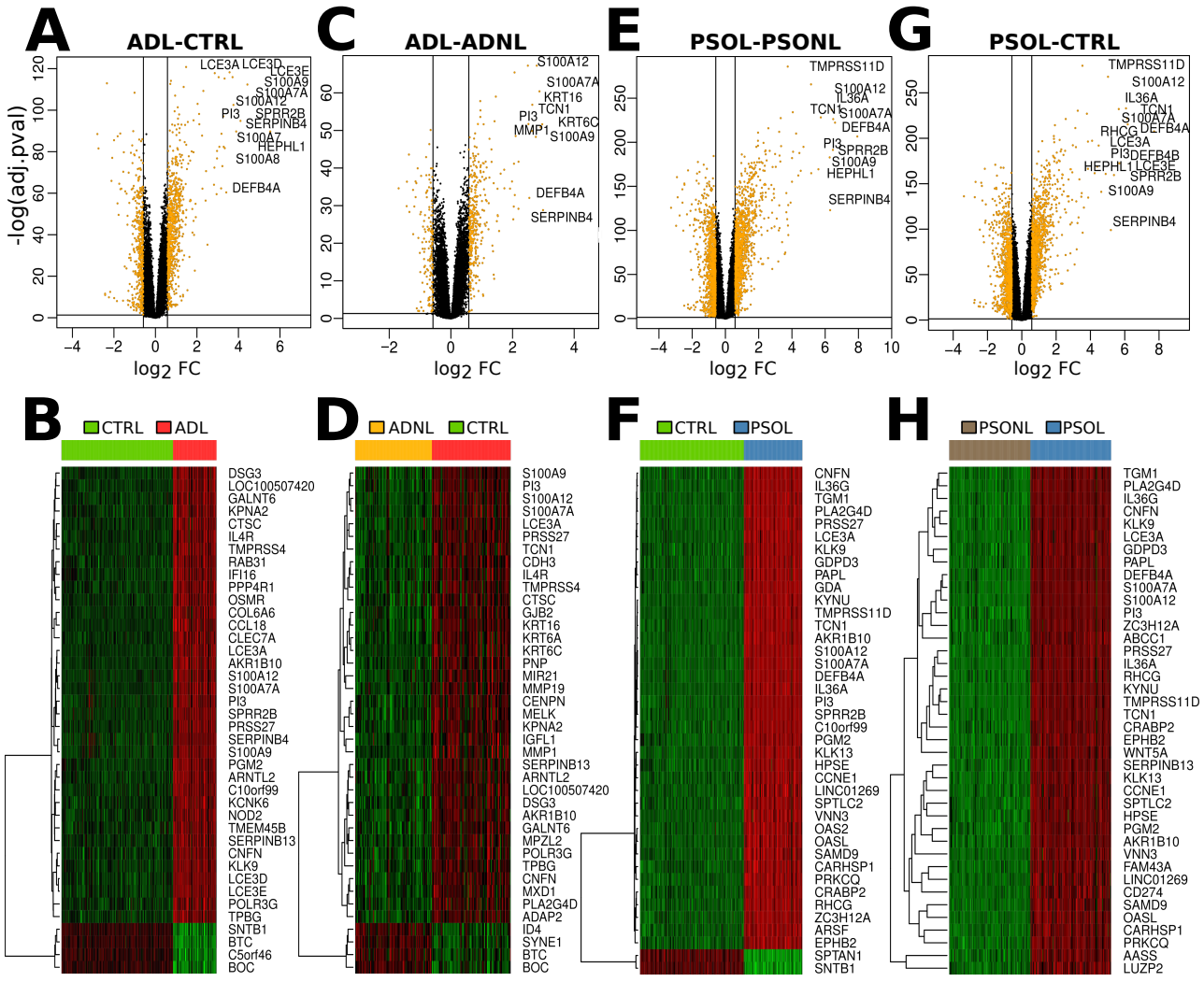

Figure 4.3: Differential expression analysis of lesional tissue. (A,C,E,G) Volcano plots. Horizontal lines correspond to p = 0.05, vertical lines correspond to log2 fold change of -0.58 and +0.58. (B,D,F,H) Heatmaps of contrasts including lesional tissue. The expression of the top 40 most significant differentially expressed genes are shown

such as LCE3D, LCE3A, LCE3E, SPRR2B and CNFN [188, 61]. Keratinocyte secreted antimicrobial peptides were differentially expressed which are known to be associated with AD [58, 188] including S100A7A, S100A9 and S100A12. Genes associated with immune response were up-regulated including IL4R, NOD2, CLEC7A and CCL18 [61, 58]. The interferon inducible gene, IFI16 was also found amongst the top 40 up-regulated genes. ARNTL2 was highly ranked and was also expressed in non-lesional tissue suggesting a potential role of the Aryl hydrocarbon receptor signalling pathway (AHR) in both uninvolved and lesional atopic dermatitis. The matrix metalloproteinase MMP12 was differentially expressed which is involved in ECM remodelling and is known marker of inflammation expressed in atopic skin [61]. These results were mirrored in the comparison of lesional atopic skin to non-lesional atopic skin which revealed 500 DEGs (353 up, 147 down, **Table 4.2**), 97% of which were also differentially expressed in ADL-CTRL (**Figure 4.3 C,D**). Several of the most highly DE genes were also amongst the top genes in lesional disease including antimicrobial peptides of the S100 family, IL4R and epidermal barrier associated genes LCE3A, and CNFN.

Pathway analysis of the most up-regulated genes in both ADL-CTRL and and ADL-ADNL revealed a set of enriched pathways that was highly concordant (**Figure 4.4 A)**. The most significant pathways included Atherosclerosis signalling pathway (ADL-CTRL p = 5.01e-15) which is associated with vascular inflammation and has recently been associated with AD [58]. Most of the top enriched pathways corresponded to the immune system and included Granulocyte Adhesion and Diapedesis (ADL-CTRL, p = 1.56e-14), T Helper cell differentiation (p = 1.58e-08) and iCOS-iCOSL signalling in T helper cells (p = 8.7e-08) which is associated with the activation of Th1 and Th2 cells. Nine genes were differentially expressed between ADL-CTRL which were involved in the 'complement system' (ADL-CTRL, p = 1.2e-04) which has not been shown in previous analysis of the atopic transcriptome suggesting the classical complement system may also contribute towards atopic inflammation. A significant enrichment for Hepatic Fibrosis / Hepatic Stellate cell activation (p = 2.29e-07) was found indicating that genes involved in fibrotic processes are up-regulated in lesions [48].

4.3.3.2 Genes and pathways upregulated in lesional psoriatic skin

A total of 2516 (1326 up, 1190 down) genes were significantly differentially expressed in PSOL samples compared to healthy with a mean log fold change of 1.0 (**Table 4.2**). The

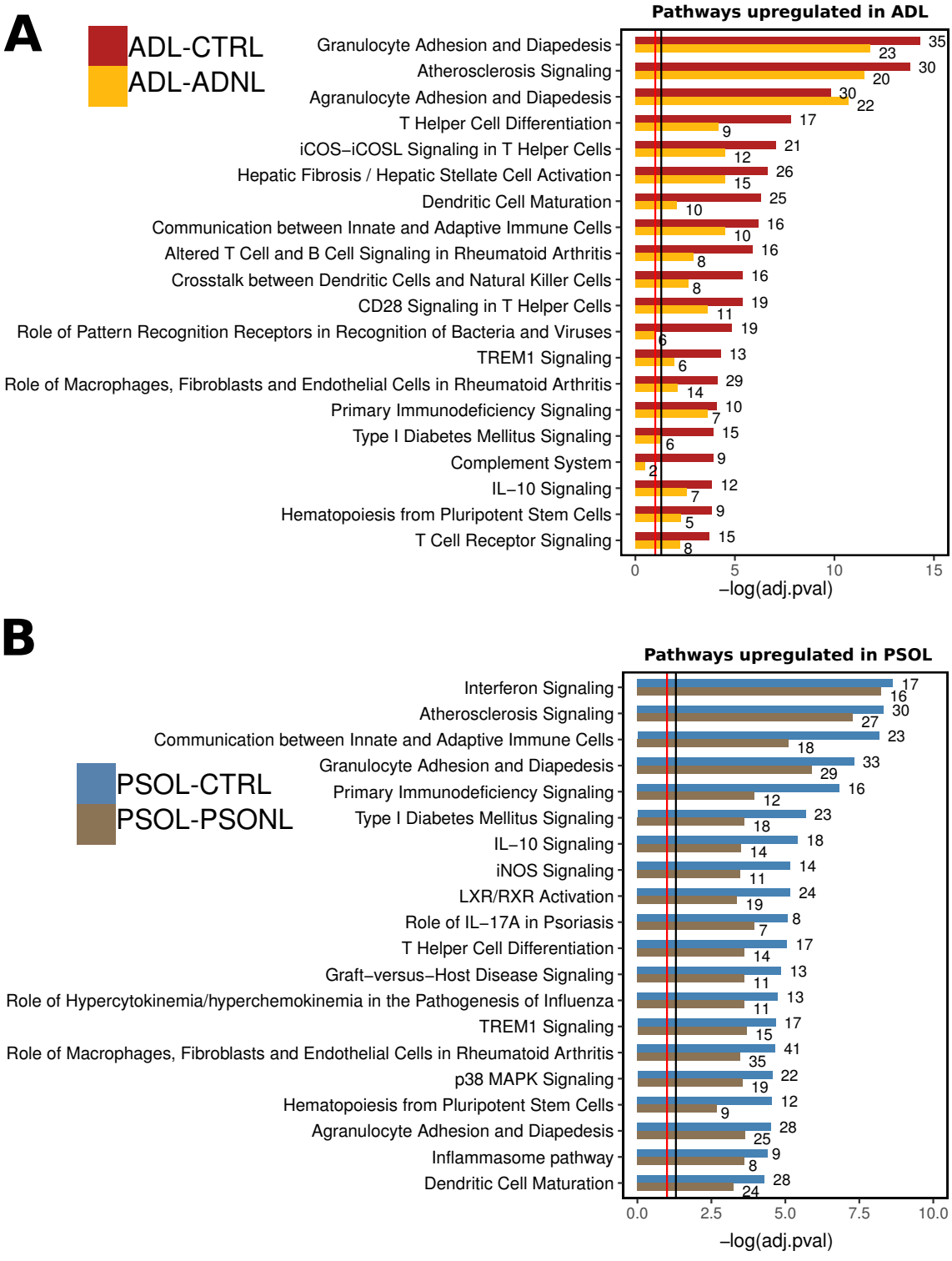

Figure 4.4: Enriched Ingenuity pathways upregulated in disease for contrasts involving lesional disease. (A) Top upregulated AD pathways enriched in ADL-CTRL (red) and ADL-CTRL (gold) contrasts. (B) Top upregulated PSO pathways enriched in PSOL-CTRL (blue) and PSOL-PSONL (gray) contrasts. All pathways shown are BH adjusted p value < 0.1 in at least one contrast. Red and black lines correspond to p = 0.1 and 0.05 respectively.

top 40 genes revealed a picture of transcriptional dysregulation in concordance with several published transcriptomes (**Figure 4.3 E,F**). Strikingly high expression levels of antimicrobial peptide genes were observed including S100A12, S100A7A and DEFB4A which were all up-regulated with $\log_2$ fold changes > 6 [77, 177, 179, 180]. Genes encoding for components of the epidermal barrier were upregulated including CNFN, TGM1, LCE3A, and SPRR2B [185, 188, 179].

As well as epidermal barrier and antimicrobial peptides; genes associated with inflammation and the immune system were up-regulated. IL36G and IL36A were expressed to extreme levels in psoriatic skin (LFC > 4) [185, 187, 179]. Other upregulated immune genes included OASL, OAS2 and TCN1. Also amongst the top genes were those associated with energy and lipid metabolism [176] including PLA2G4D [179], KYNU [140], GDPD3 and RHCG. As observed in AD, the differentially expressed genes between PSOL-PSONL closely resembled those identified in PSOL-CTRL (**Figure 4.3 G,H**) including IL36, antimicrobial peptides, EDC genes and those corresponding to lipid metabolism.

Interferon signalling was the top enriched pathway amongst genes upregulated in PSOL ($p = 2.34e\text{-}09$, **Figure 4.4 B**). Several other immune pathways were enriched including 'Communication between innate and adaptive immune cells' (PSOL-CTRL; $p = 6.76e\text{-}09$), which contained TLR2 and was specifically expressed in PSO. 'iNOS signalling' ($p = 6.76e\text{-}06$) contained the highly up-regulated NOS2 (LFC = 2.8) and indicates activation of macrophages, 'TREM1 signalling' ($p = 2.13e\text{-}05$) which was recently identified in an AD transcriptome study [173] and 'IL10 signalling' ($p = 3.81e\text{-}06$). Significant enrichment for 'LXR/RXR activation' ($p = 6.76e\text{-}06$) was found and included apolipoprotein 1 (APOL1) suggesting potential disruption of processes involved in lipid and cholesterol biosynthesis.

4.3.4 Genes and pathways downregulated in lesional skin

4.3.4.1 Genes and pathways downregulated in lesional atopic skin

Differential analysis revealed 463 genes downregulated in ADL compared to healthy individuals (**Table 4.2**). Few pathways were enriched amongst down regulated genes (**Figure 4.5 A)**. The top pathway was 'Oleate Biosynthesis II' (ADL-CTRL; p = 0.009) and contained the genes FADS1, FADS2, SCD5 and ALDH6A1. These results reflect previous observations of abnormal lipid composition in the stratum corneum of patients with AD [59]. Further enrichment was observed in 'wnt-catenin signalling', establishing a link between this pathway in atopic inflammation. Five genes involved in circadian rhythm signalling (ADL-CTRL; p = 0.03) were identified including PER3, PER1, NR1D1, BHLHE41, CRY2. Circadian rhythms have recently been shown to control the expression of proinflammatory cytokines [191] thus this pathway could play a role in atopic inflammation. As coverage of enriched pathways was poor amongst downregulated pathways, enrichment of GO [142] biological processes with enrichR [189] was performed (**Figure 4.5 B)**. This analysis revealed enrichment amongst several processes involved in fatty acid and lipid metabolic processes improving support for a transcriptional profile associated with disruption to lipid functionality. Circadian regulation of gene expression was also enriched providing further support for disruption of this process in AD.

4.3.4.2 Genes and pathways downregulated in lesional psoriatic skin

Comparison of lesional PSO to healthy revealed no Ingenuity pathways which were enriched ($p < 0.1$) however several pathways were enriched in the PSOL-PSONL contrast. Enriched pathways were associated with lipid biosynthesis including (**Figure 4.5 C**) 'LPS/IL1 Mediated inhibition of RXR function' (PSOL-PSONL; p = 0.033) and 'Oleate Biosynthesis II' (PSOL-PSONL; p = 0.044). The same four genes involved in 'Oleate Biosynthesis II' found in AD, were also downregulated in PSO suggesting lipid biosynthesis is also dysfunctional in psoriatic tissue. Pathways relating to energy metabolism including 'AMPK signalling' (PSOL-PSONL; p = 0.03) and 'leptin signalling in Obesity' (PSOL-PSONL; p = 0.03) were significantly down and was further supported by GO enrichment of energy metabolic processes such as 'regulation of glucose', 'carbohydrate metabolic process' and 'fatty acid metabolic process' (**Figure 4.5 D**). 'Axonal guidance signalling' (PSOL-PSONL; p = 0.08) was enriched amongst downregulated genes. It been suggested that neurons and neuropeptides may be associated with pruritus in PSO [192, 193]. Mild to severe pruritus

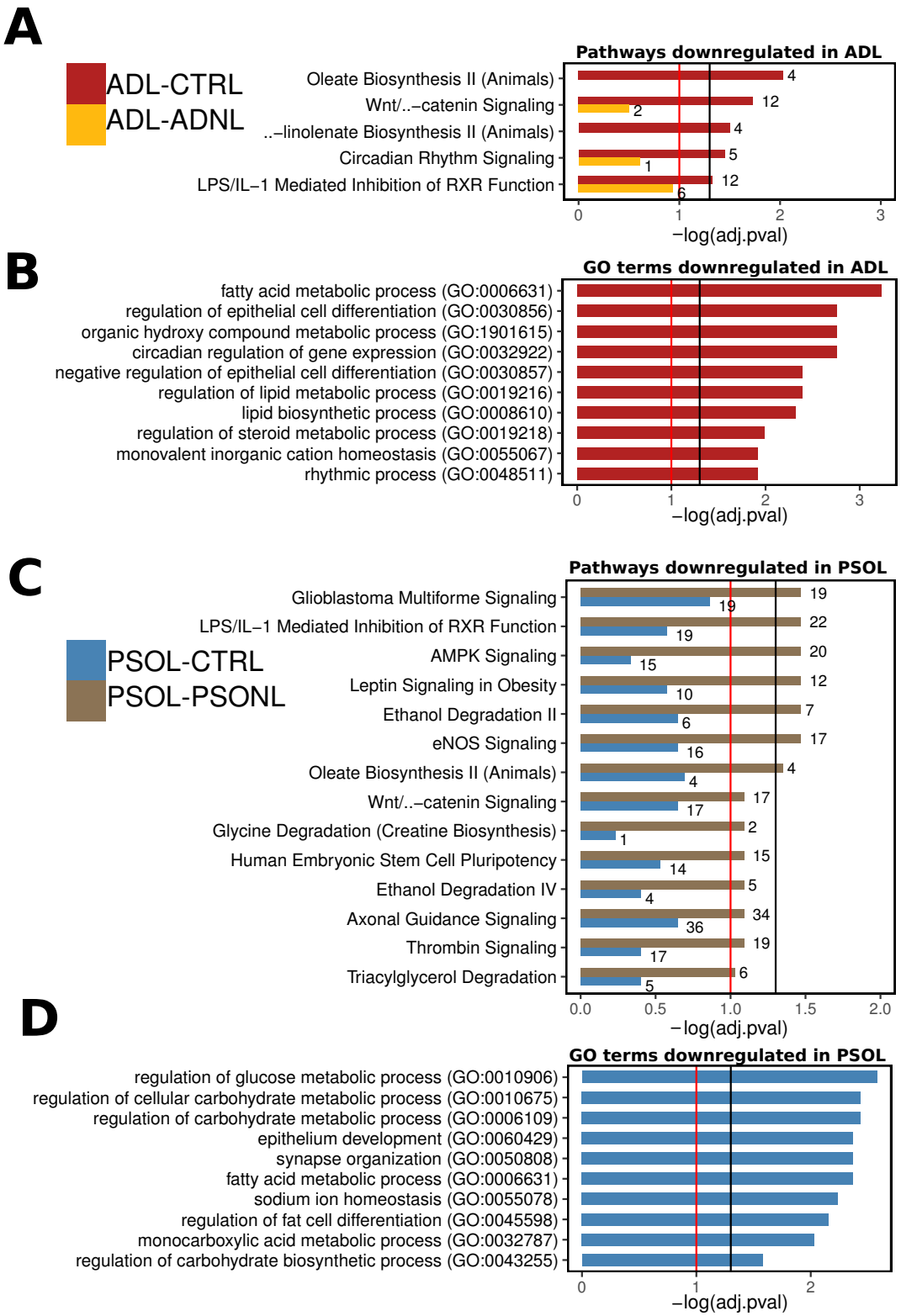

Figure 4.5: Enriched Ingenuity pathways downregulated in disease for contrasts involving lesional disease. (A) Top downregulated AD pathways enriched in ADL-CTRL (red) and ADL-CTRL (gold) contrasts. (B) Top enriched GO terms amongst downregulated genes in ADL-CTRL. (C) Top downregulated PSO pathways enriched in PSOL-CTRL (blue) and PSOL-PSONL (gray) contrasts. (D) Top enriched GO terms amongst downregulated genes in PSOL-CTRL. All pathways shown are BH adjusted p value < 0.1 in at least one contrast. Red and black lines correspond to p $= 0.1$ and 0.05 respectively

affects approximately 60-90% of psoriasis patients [194]. An increase in nerve growth factor (NGF) with ELISA has been found in psoriatic individuals with pruritus [195] and whilst NGF was not found to be differentially expressed (in either AD or PSO), 34 down-regulated genes associated with neuronal guidance were down-regulated in psoriatic lesions. These included members of the Ephrin family and receptors, EFNB2 and EPHB1, Semaphorins including, SEMA3D and SEMA3E, Plexins including PLXNA3, and SLIT2. Further support was identified by enrichment for the 'synapse organisation' GO term (**Figure 4.5 D**) and potentially relate to a neurological axis which relate to the mechanisms of pruritus. 'Wnt β-catenin signalling' was found to down-regulated (PSOL-PSONL; p = 0.08) and supports previous reports of dysfunctional Wnt signalling in psoriasis [190, 181].

4.4 Disease specific gene sets

To gain insight into transcriptional processes that are common between diseases, and those which are preferentially expressed in either AD or PSO, signatures were defined by intersecting the lists of differentially expressed genes (**Figure 4.6**). To evaluate differences in immune activation, cytokines and chemokines preferentially up-regulated in AD or PSO were identified (**Figure 4.7**).

4.4.1 Common and disease associated inflammatory signatures

4.4.1.1 Common inflammatory gene signatures and pathways

Six hundred and forty one genes were commonly up-regulated between both diseases (**Figure 4.6 A**). The most significant pathways included Atherosclerosis Signalling (p = 1.99e-08) which was identified in a recent AD meta-analysis [58], Granulocyte Adhesion and Diapedisis (p = 3.63e-07) and T helper cell differentiation (p = 5.62e-06). To determine core similarities and differences in immune response, cytokine and chemokine signatures of the commonly upregulated genes were identified (**Figure 4.7**) using the same DEG sub-setting approach employed by Guaranta et al. [185]. IL22, IL36A and IL36G were up-regulated in both diseases, as well as an array of cytokine receptors. These included IL36RN, IL4R, IL7R, IL10RA and IL2RA. A chemokine profile of CXC family CXCL8, CXCL9, CXCL10, CXCL11, CXCL17, and CC family CCL2, CCL22, CCL18 and CCL19 were up-regulated in both diseases. The pathways, as well as proinflammatory genes identified reflect a core disease signature associated with vascular inflammation, immune cell

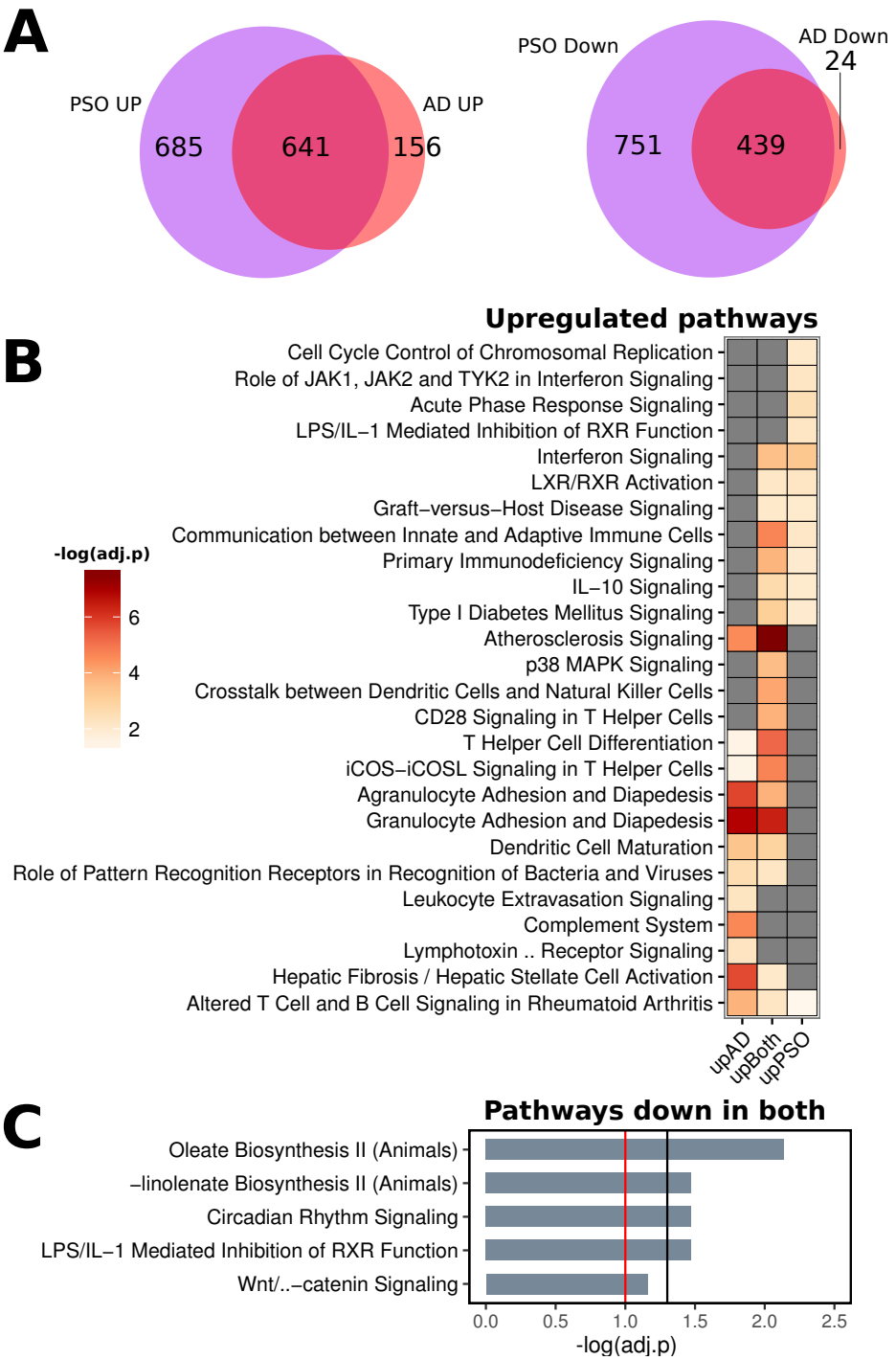

Figure 4.6: Cohort specific gene sets and pathway analysis. (A) Venn diagrams showing overlap of upregulated and downregulated genes between ADL and PSOL compared to control. (B) Top significant pathways for up-regulated gene sets. Heatmap tile colors represent -log of the adjusted p value for pathway enrichment. Gray tiles are not significant. (C) Enriched pathways amongst genes downregulated in both ADL and PSOL.

chemotaxis and lymphocyte activation (**Figure 4.6 B**). It is important to note that for many of commonly upregulated immune genes, the expression in PSO was higher than that in AD except in the cases of: IL2RG, IL2RA, IL10RA, CCR4, CCR1 and the CC chemokines CCL2, CCL18, and CCL19.

4.4.1.2 Genes and pathways preferentially expressed in AD

To identify biological processes associated with a specific disease, pathway enrichment of gene sets preferentially expressed in in either AD or PSO was performed. Despite a high degree of overlap, several pathways were enriched within PSO and AD specific DEGs (**Figure 4.6 B**). 156 genes were specifically up regulated in atopic lesions ($p < 0.05$, LFC > 0.58). The chemokine and cytokine profile of AD specific genes was evaluated (**Figure 4.7**) revealing a Th2 signature including cytokines IL10 and IL24, as well as IL13RA2 and the chemokine receptor CCR8. TGFB was found to be of increased expression in AD along with a panel of CC family chemokines associated with immune cell migration including CCL17, CCL8, CCL26, CCL1 and CCL13. With upstream regulator analysis, the Th2 cytokine IL4 ($p = 1.67e\text{-}28$) was found to be the most significant candidate.

The most significant up-regulated pathway was 'Granulocyte Adhesion and Diapedesis' ($p = 1.25e\text{-}07$) which is involved in migration of immune cells and contains several CC family chemokines (as described above) indicating differences in immune cell chemoattraction between AD and PSO (**Figure 4.7**). These findings support previous observations that AD is associated with over expression of CC family, and PSO with CXC family chemokines [183]. This pathway also included VCAM1 which is a cell adhesion molecule induced by Th2 cytokines [196].

Hepatic Fibrosis / Hepatic Stellate Cell activation' ($p = 1.86e\text{-}06$) was enriched in AD specific genes and consisted of collagen transcripts, COL4A1, COL6A3, COL6A5, COL6A6, the TIMP metallopeptidase 1 (TIMP1) and immune genes IL10 and TGFB1. TGFB was expressed specifically in ADL and is known to be pro-fibrotic. Furthermore, it has been suggested that extracellular matrix remodelling in AD could be driven by TGFB [197]. These results indicate that ECM remodelling may be associated with skin thickening that accompanies AD [198]. As TGFB was preferentially expressed in AD, it could be that TGFB is a differentiator which drives this response. AD specific genes were also strongly enriched for complement system ($p = 2.45e\text{-}05$) establishing a link between this pathway

and atopic inflammation.

4.4.1.3 Genes and pathways preferentially expressed in PSO

The psoriatic transcriptome was characterised by 685 specifically up-regulated genes (**Figure 4.6 A**). Evaluation of cytokine and chemokine profiles identified an up-regulated Th17 signature including IL17A, STAT3 and CCL20 (**Figure 4.7**). The Th1 cytokine IFNG was up-regulated in PSO, as well as high expression of interferon inducible gene IFIT1. Several proinflammatory cytokines including IL19, IL20 and IL33 were up-regulated in PSO, as well as a panel of chemokines mediating immune cell trafficking; these consist of CXCL1, CXCL16, CXCL13 and CCL20. IL1B was specifically expressed in psoriasis samples. Whilst the IL1 family cytokines IL36G and IL36A were differentially expressed in both AD and PSO, they were amongst the most up-regulated genes in PSO with log fold changes > 4, therefore, these results confirm recent reports emphasising the importance of this family in the pathogenesis of psoriasis [199, 179]. Top upstream regulators of up-regulated PSO genes were INFG (p = 6.12e-24) and STAT3 (p = 9.16e-24).

Pathway enrichment amongst genes preferentially expressed in PSO revealed Th1 and psoriasis associated pathways such as 'Interferon Signalling' and 'Role of JAK1, JAK2, and TYK2 in Interferon signalling' (**Figure 4.6 B**). 'Interferon signalling' contained INFG which is produced by Th1 cells and is considered to be one of the main differentiators between AD and PSO [184]. Psoriasis associated genes were enriched for 'Acute phase response signalling' (p = 0.003, **Figure 4.6 B**). This pathway included an array of immune transcripts including the IL1 family cytokines IL1B, IL36B, and IL33, together with the receptor IL1RN. The authors of [200] found up-regulated serum levels of IL33 in AD compared to psoriasis, however, here the opposite trend was observed in the skin transcriptome (AD LFC = -0.04, PSO LFC = 0.72). Other genes in this pathway consisted of STAT3 and NFKBIB, all of which were preferentially up-regulated in PSO.

Genes involved in 'LPS/IL-1 Mediated inhibition of RXR function' were over-represented (p = 0.006) and contained IL1 family cytokines and genes associated with lipid and xenobiotic metabolism, as well as enrichment for 'LXR/RXR Activation' (p = 0.006) which included NOS2, a known biomarker in PSO [201], thus, enrichment of this pathway may correspond to the metabolic syndrome which accompanies patients with psoriatic lesions [72].

Table 4.4: Genes expressed in opposite directions

Gene	LFC AD-CTRL	P AD-CTRL	LFC PSO-CTRL	P PSO-CTRL	Description
CEACAM5	-0.62	3.0e-09	0.77	9.4e-16	Carcinoembryonic Antigen
ASPN	0.60	4.9e-17	-0.83	8.6e-36	Asporin
PPP1R3C	0.69	2.6e-22	-0.86	5.09e-38	Protein Phosphatase subunit

4.4.1.4 Commonly downregulated pathways

There were 439 down-regulated genes in both AD and PSO (**Figure 4.6 A**). These genes were associated with lipid biosynthesis pathways (**Figure 4.6 C**) including 'Oleate Biosynthesis' ($p = 0.007$) and 'LPS/IL-1 Mediated inhibition of RXR function' ($p = 0.03$). The genes involved in 'Oleate Biosynthesis' were SCD5, FADS2, FADS1 and ALDH6A1 indicating that lipid biosynthesis is dysfunctional in both diseases. 'Circadian rhythm signalling' was significantly downregulated in both diseases ($p = 0.03$) which has recently been shown to regulate expression of proinflammatory cytokines [191] therefore, this finding could establish a general link with skin inflammation. Whilst no strong enrichment of immune system processes was found, cytokines of reduced expression in both diseases included IL34 and IL37.

4.4.2 Genes expressed in opposite directions

Genes were filtered for those which were differentially expressed in opposite directions i.e., genes which were significantly up-regulated in AD ($p < 0.05$, LFC > 0.58), however, were significantly down-regulated in PSO ($p < 0.05$, LFC < -0.58, **Table 4.4**). Only three genes satisfied these criteria including CEACAM5, which is involved in cell adhesion, Asporin (ASPN) which was up-regulated in AD but down regulated in PSO, is an ECM protein which is capable of inhibiting TGFB signalling in cartilage [202]. ASPN was also found to be significantly up-regulated in non-lesional atopic skin and therefore could be an interesting candidate for further study. PPP1R3C is involved with protein phosphorylation with a wide range of functions. The function of these genes in the skin is currently unknown, thus further investigation may provide insights into the mechanisms which differentiate disease.

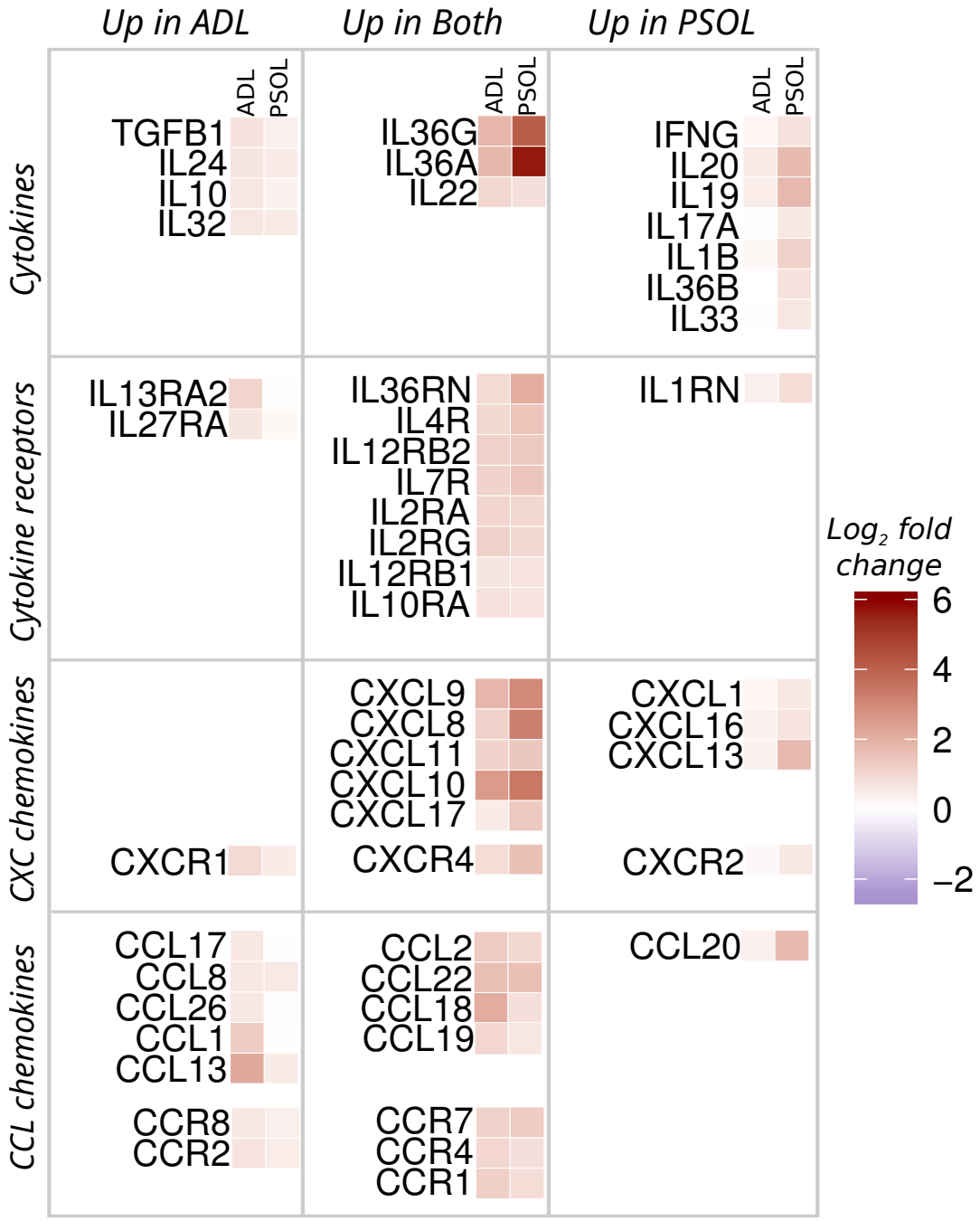

Figure 4.7: Common and specific cytokine expression. Up in ADL corresponds to DEGs which were LFC > 0.58, p > 0.05 in ADL-CTRL and not in PSOL-CTRL.

4.5 Conclusions and Discussion

PCA showed that lesional samples had distinct transcriptomes from uninvolved and healthy skin, and that the major axis of variability was associated with immune response. Furthermore, non-lesional samples were almost indistinguishable to healthy controls. This was supported by the finding that PSONL had very few DEGs (26) suggesting that non-inflamed skin closely resembles healthy tissue. On the other hand, transcriptional dysregulation in lesional skin was vastly perturbed with over 2500 genes. A different pattern emerged in AD as uninvolved tissue had 86 genes of altered expression whereas lesions were perturbed to a smaller extent than PSO. Atopic lesions were characterised by fewer DEGs than psoriasis which is in concordance with other comparative studies [185, 183]. The fold changes of uninvolved DEGs tended to be higher in AD; however, in a lesional phase, the expression in PSO was greater than in AD suggesting that susceptible AD skin exists in a heightened state of disorder, however during a flare, transcriptional dysregulation in PSO overshadows AD.

4.5.1 Uninvolved skin

Epidermal barrier dysfunction is characteristic of skin susceptible to disease
Considering the non-lesional tissue of both diseases, most DEGs in PSONL were also significant in ADNL. Amongst the overlapping genes, those encoding components of the epidermal barrier including SPRR2G, SPRR2B along with the antimicrobial peptides S100A9 and S100A7 were identified. These genes were significantly up-regulated clearly suggesting that barrier dysfunction precedes a flare event.

Heightened immune activation in non-lesional AD
An overlapping immune signature within uninvolved skin was identified which included up-regulation of interferon inducible genes CXCL10, CXCL9, and IFI27. These genes represent a component of the immune system which is activated in the absence of inflammation of both diseases. Further immune system activation was observed in non-lesional AD with the specific up-regulation of the chemokines CCL18 and CCL13. CXCL10 and CCL18 have been identified in the non-lesional skin of AD patients [61], however, to the best of our knowledge CXCL9 and CCL13 have not been reported. The expression of CCL18 is enhanced by the expression of Th2 cytokines [203], therefore, it could represent a Th2 signature within the uninvolved skin of AD and indicates that non-lesional skin exists in

a state of heightened immune activation. Taken together, this analysis indicates that susceptible AD skin is associated with Th1/Th2 immune activation, increased antimicrobial peptide presence and increased expression of genes involved in terminal differentiation. Uninvolved psoriasis more closely resembled healthy skin, but displayed evidence of increased Th1 activity and dysregulation amongst genes encoding for the epidermal barrier.

4.5.2 Lesional skin

The epidermal barrier is significantly disrupted during inflammation

Both AD and PSO were associated with transcriptional changes to integral components of the epidermal barrier. Several members of the LCE3 family along with members of the SPRR family were up-regulated in both diseases as well as CNFN. Disruption to barrier gene expression has been identified in previous studies [185, 188] and is a major area of interest in AD as barrier weakness may convey increased susceptibility to transepidermal allergen transfer [204, 205] and colonisation of pathogens like *Staphylococcus aureus* [206]. Interestingly, disruption to the barrier in PSO was potentially greater as fold changes were more extreme, therefore, barrier weakness alone does not explain AD. It is also possible that increased expression of epidermal genes in PSO is reflective of increased proportion of keratinocytes in psoriatic biopsies due to hyperproliferation [207].

Common immune system dysregulation

The immune system was vastly perturbed in both diseases as shown by the common up-regulated genes. A common inflammatory signature was enriched for immune processes including 'Granulocyte adhesion and diapedesis' and 'Atherosclerosis signalling'. This common immune signature consisted of CCL22, CCL2, and CXC family chemokines as well as IL22, ICOS, IL36A and IL36G. Whilst IL36 was differentially expressed in both diseases, the expression of IL36 was at strikingly high levels in PSO, as others have shown [179]. The common signature also included several cytokine receptors including IL36RN, IL10RA, IL2RA, IL7R and IL4R. The top upstream regulator was IFNG (p=1.8e-46), therefore, the common inflammatory immune signature is likely to be associated with Th1 activity which is characteristic of PSO [73], and also characteristic of AD in the chronic phase [44].

Many antimicrobial peptides were up-regulated in both diseases and including S100A7A, S100A7, S100A2, S100A12, S100A8, S100A9 and DEFB4A. Whilst AMPs were dysregulated in both, they tended to be of greater expression in PSO as others have found [77].

Th2 activation in AD and Th1/Th17 activity in PSO

The underlying immunological consensus places AD as a disease associated with a Th2 cell response in the acute phase [44], whereas PSO is thought to be associated with interferon-γ producing Th1 cells and IL-17A producing Th17 cells [73]. The cytokine profile preferentially expressed in AD, included IL10, IL24 and IL13RA2 along with an array of CC-family cytokines and CCR8. Upstream regulator analysis identified IL4 as the top candidate amongst genes preferentially expressed in AD, therefore, these findings support the hypothesis of Th2 disbalance in AD.

In psoriasis, preferential up-regulation of IL1B, IL19, IL20, IL33, INFG and IL17A was found along with an array of CXC family cytokines. Pathways associated with PSO specific genes were classic Th1 pathways such as 'Interferon signalling'. Top upstream regulators were INFG and STAT3, thus, these results suggest that PSO is of higher expression in Th1/Th17 cytokines and inducible products, and supports the current consensus model of these diseases.

Disruption to lipid biosynthesis in AD and PSO

Disruption to lipid biosynthesis has been associated with both AD and PSO previously [58, 140]. Pathways involved in lipid biosynthesis including 'Oleate biosynthesis II' were found to be enriched amongst down-regulated genes in AD compared to healthy samples supporting observations of reduced lipids and ceramides in AD [60, 59]. As a larger number of downregulated genes were found in PSO (1190 vs 463), this pathway was not found to be significant, however, the same genes were dysregulated suggesting that this characteristic is also likely to be present in psoriatic skin.

Further analysis of GO terms showed a downregulation of genes involved in fatty acid metabolism for both diseases. Within the epidermis, keratinocytes synthesise lipids which are a necessary component of the cornified layer and as lipid bioysnthesis has been linked to barrier dysfunction, this could also be associated with the colonisation of *Staphylococcus aureus* [206]. Wnt signalling is known to be disrupted in psoriasis [181, 190] although downregulated genes were enriched in both diseases establishing a general link between this pathway and skin inflammation. Core genes involved in the circadian clock were also found to be downregulated in both diseases. Expression of certain cytokines have been shown to

be under clock control, and mutations within clock inhibitors can result in a psoriasiform inflammation in mice [191].

High expression of ECM genes in AD
AD was characterised by high expression of extracellular matrix genes and was enriched for pathways such as a Hepatic 'Fibrosis/Hepatic Stellate Cell Activation'. Several members of the matrix metalloproteinases family; MMP3, MMP19, TIMP1, as well as members of the collagen family; COL4A1, COL6A3, COL6A5, COL6A6, were preferentially up-regulated in AD. On the other hand, only weak enrichment for ECM terms was found amongst all upregulated PSO genes compared to healthy ($p = 0.06$, pathway rank = 90). Th2 cytokines as well as TGFB were preferentially expressed in ADL and can induce fibrosis [198, 197, 208] which may explain the differential expression of these genes.

Down-regulation of axonal guidance signalling and energy metabolism in PSO
Genes downregulated in PSOL were enriched for 'axonal guidance signalling' which was also identified by GO enrichment analysis. Neurological factors are thought to be associated with pruritus [192, 195] in the skin, therefore, this pathway may correspond to differences in itch mechanisms. Furthermore, these findings could relate to the recent association between a set of nociceptors required to drive IL-23-mediated psoriasiform inflammation in mice [209]. Genes downregulated in PSO were also enriched for 'AMPK signalling' and 'glucose metabolism' which reflects perturbation to energy metabolism and could be associated with the metabolic syndrome which is often associated with psoriasis [72].

Lastly, the increased power due to the size of our cohort allowed us to find three genes (CEACAM5, ASPN and PPP1R3C) which were differentially expressed in opposite directions between AD and PSO. To the best of our knowledge, this is the first time these genes have been reported. Due to unique expression profiles of these genes, they are candidates for further investigation as their functional roles in the skin are not well characterised.

Overall, this analysis presents a comprehensive overview of the transcriptional differences between both lesional and uninvolved skin from AD and PSO. Common transcriptomic components were identified, and signatures which were preferentially expressed in AD and PSO were established. This analysis has confirmed several dysregulated processes which have been identified in previous studies such as Th1/Th17 and Th2 polarisation in PSO

and AD respectively, as well as disruption to the epidermal barrier and lipid bioysynthesis pathways. Several other components which are less established and have not been reported in previous transcriptome analysis were identified such as disruption to the ECM, neurological components and circadian rhythms. This transcriptome wide exploratory analysis places AD as a disease characteristic of heightened Th2 immunity, increased ECM remodelling and with disruption to the epidermal barrier. On the other hand, PSO was associated with increased activity of Th1/Th17 immunity, extreme antimicrobial peptide expression, abnormal epidermal barrier expression and with potential disruption to energy metabolism and axonal guidance signalling pathways. Both diseases showed traits of disrupted lipid and fatty acid metabolism. The gene sets identified in this chapter characterise the major differences and establish a basis for integration with the cutaneous microbiome.

Chapter 5

Host-microbe integration

5.1 Introduction

It is now generally accepted that the resident microbiota plays a role in shaping the immune system and helps with the maintenance of homeostatic equilibrium [22, 21]. When the delicate balance in the microbiota is disrupted, the composition of microbial communities enters a state of dysbiosis which has been linked with chronic inflammatory diseases [87] such as Crohn's disease [4, 5], atopic dermatitis [64] and asthma [120]. If dysbiosis does indeed exacerbate or trigger inflammation, establishing the links between resident microbiota and host response is a critical step towards development of therapeutics which can modulate the commensal microbiota.

Several previous studies have attempted to integrate the microbiome with host derived parameters such as the metabolome and transcriptome. These include Schwartz et al. [116], who used a strategy based upon canonical correlation analysis (CCA) to express a relationship between metagenomic virulence factors and intestinal host-immune gene expression in the context of diet amongst formula and breast-fed infants. Investigation of host-microbe associations with linear modelling is popular due to its ability to control for extraneous sources of variation such as age, antibiotic usage, gender and body site [33, 40, 41] which are known to impact upon microbial abundance. One such study found a trend for increased *Proteobacteria* abundance with the expression of host IL1A in the context of asthma [120]. Another study used linear models to investigate the influence of phylum level microbiota on the expression of the genes APOA1 and DUOX2 [119] revealing clinically relevant interactions in the context of Crohn's disease and ulcerative colitis.

A more recent application simultaneously reduced dimensionality in both microbiome and transcriptome datasets to maximise power, and then used linear models to identify associations between host transcriptional and microbiomic factors in the inflamed ileal pouch [29].

In previous chapters, the overall properties of resident community composition, as well as the host transcriptional architecture of skin inflammation were determined. Given that the MAARS consortium data consists of both microbiome and transcriptome resources sampled at the site of inflammatory disease, this provided a unique opportunity to study host-microbe interactions. This chapter presents an exploratory analysis into microbe-associated host-transcriptional pathways. To investigate host-microbiome associations, a power analysis was performed to approximate an appropriate number of possible hypothesis tests given an assumed covariance of 0.5 between host transcripts and taxa whilst retaining 80% power [29]. Next, dimensionality reduction schemes were performed to simultaneously reduce the number of transcripts and species subjected to hypothesis testing.

Several strategies were implemented to detect host-microbe interactions. In an unbiased analysis to identify global patterns, abundant taxa and disease associated genes (as identified in Chapter 4) were hierarchically clustered and then linear models were implemented to identify relationships between gene clusters and taxa clusters [29]. Targeted approaches for identifying transcriptomic patterns associated with candidate OTUs were also implemented. For key species such as *S. aureus* in AD and *C. simulans* in PSO, patients were stratified into discrete groups based upon median OTU abundance, and then differential analysis of the transcriptome using limma [138] was performed to identify pathogen associated transcriptomic signatures. In an alternative approach, host transcriptomic factors were derived by principal component analysis and pairwise linear models were applied comparing the relative abundance of selected species with the set of principal components (PCs) that explained 50% of the total variance in the transcriptome [29].

Initial discovery of the *S. aureus* associated gene signature via dichotomisation of the host transcriptome was performed by Marine Jeanmougin. A reimplementation of this analysis is presented in **Figure 5.7** and has been extended to determine the association with disease severity.

5.2 Methods

5.2.1 Sample selection

Quality controlled and RMA normalised gene expression data, was obtained from Institut Curie, and quality controlled and normalised 16S microbiome sequencing data was obtained from the Karolinska institutet and Institut Curie as part of the MAARS consortium project. Details of patient recruitment, sampling and data processing are detailed in **(Sections 2.3.1, 2.3.2 and 2.3.3)**. For integrative analysis, only samples with both microbiome and transcriptome samples were selected as described in (**Table 5.1**).

5.2.2 Power analysis

To guide the scale of dimensionality reduction required, a power analysis was performed under the framework described by Morgan et al. [29]. It was used to determine the number of possible pairwise tests to detect a true covariance of 0.5, whilst retaining 80% power given the available 82 ADL and 119 PSOL samples. Ten thousand correlated variable pairs of length 82 in ADL and 119 in PSOL were sampled from a bivariate normal distribution with covariance ranging from 0-1 using the R function *rmvnorm*. The 80th percentile of p values for each value of covariance was calculated representing 80% of the p values at given covariance. Then, to account for the family wise error rate (FWER) at an alpha level of 0.05, the estimated number of possible pairwise tests was calculated as 0.05 divided by the 80th percentile of p values [29].

5.2.3 Transcriptome dimensionality reduction

First, the 32633 genes present on the array were filtered to those that were associated with disease by performing differential analysis between healthy and diseased cohorts using the limma package [138] as described in **Chapter 4** and **Section 4.2.3**. This initial filtering resulted in a total of 1260 AD and 2516 PSO associated genes ($p < 0.05$, LFC > 0.58). DEGs were then split into upregulated and downregulated groups, and were independently clustered with hierarchical clustering using 1 - Pearson correlation distance and Ward's linkage method. To further reduce DEGs, the optimal number of clusters in each set was estimated. For clusters, $2 \leq k \leq 50$, the average silhouette width was calculated, and k was selected as the cluster with the maximum silhouette width. A single representative of the differential gene cluster was calculated as the cluster centroid, i.e., the

average expression of genes within the differential gene cluster (**Figure 5.2**). Calculation of gene centroids have been performed in a similar way by Kivela et al. [210]. In a parallel analysis, the cluster representative was calculated as the first principal component where the same conclusions were reached.

In an additional strategy to perform host transcriptome dimensionality reduction, genes were filtered to those with at least the median variance across the transcriptome resulting in 16316 genes. Principal component analysis of the scaled and centred transcript expression was performed and the number of components which explained 50% of the variance were retained [29].

5.2.4 Microbiome dimensionality reduction

To select taxa for integration, all OTUs were summed up at each taxonomic level and minimum abundance filtering was performed to remove rare species. Each taxon was required to be present in at least 20% of samples with a mean relative abundance of at least 0.005. Average linkage hierarchical clustering of arcsine square-root transformed relative abundance of taxa was performed using 1 - Pearson correlation distance and the dendrogram was cut at a fixed height of 0.5. A cluster representative from each was selected as the taxon with the lowest mean abundance [29] (**Figure 5.2**).

5.2.5 Host-microbe associations

To estimate host-microbe associations between dimensionality reduced microbe cluster representatives and transcript cluster centroids, pairwise linear models were applied. To account for potential confounding factors known to be associated with microbial relative abundances, the formula: gene centroid $\sim$ taxa cluster representative + body site + sampling institution + gender + age was used. Relative abundances were arcsine square root transformed [5]. Significant associations were considered as those with a Benjamini Hochberg corrected p value < 0.1 with a correlation coefficient > 0.4 (**Figure 5.2**). The same model formula was used to identify associations between expression principal components and taxa of interest.

5.2.6 Microbe associated transcriptional signatures

For taxa of interest, patients were grouped into 'High' and 'Low' groups based upon median OTU abundance. Using the high and low groups, the transcriptome was dichotomised and differential analysis was performed with the limma package [138]. Differentially expressed genes between OTU-high and OTU-low were considered as those with an adjusted p value $<$ 0.1. Dichotomised taxa groups were also used to test for differences in principal component scores. Association of high and low groups with local SCORAD was performed with a Wilcoxon ranked sum test.

5.2.7 Functional analysis

Over representation of pathways was performed with Ingenuity Pathway Analysis [143]. Gene ontology enrichment was performed with enrichR [189]. Annotation of gene principal components was performed by selecting the top 20 loadings of the greatest magnitude in both positive and negative directions. A pre-ranked gene set enrichment analysis [147] was performed to attribute functions to the principal component loadings using the REACTOME database [145].

5.3 Results

To uncover potential host-microbe interactions in lesional disease, samples were first restricted to matched lesional AD and PSO samples, i.e., those which had both a microbiome and transcriptome in the MAARS cohort as described in (**Table 5.1**).

Table 5.1: Matched transcriptome microbiome integration cohort

		ADL	PSOL
Patients(n)		82	119
Samples(n)		82	119
Gender (n)	Female	36	26
	Male	46	93
Anatomical Location (n)	Buttocks	-	18
	Lower Back	2	90
	Thigh	43	1
	Upper Back	37	10
Institution (n)	HHU	34	44
	KINGS	13	41
	UH	35	34
Age	Mean	44.5	48.8
	SD	14.5	13.6

5.3.1 Power analysis

Exhaustive pairwise comparisons between taxa and host transcripts would require a total of 32633 (transcripts) * 7532 (taxa) = 2.45e8 hypothesis tests, therefore, an association would require a raw p value of <2e-10 to be declared significant after accounting for the FWER. Such levels of significance between the microbiome and host parameters are unobtainable as reflected in the results of previous integrative studies [120, 119, 29], therefore, a suitable scheme of dimensionality reduction is required to reduce the number of pairwise tests. As discussed earlier, performing a power analysis is an important step in estimating the required magnitude of dimensionality reduction [29]. Power analysis performed as described in [29], showed that for a true covariance of 0.5 and to retain 80% power, approximately 230 tests in ADL for a sample size of 82, and 10000 in PSOL for a sample size of 119 could be performed (**Figure 5.1**). Given the disparity between the estimated number of possible tests between ADL and PSOL, the exploratory analysis was designed to accommodate

the constraints of ADL, where the objective was to reduce the the number of taxa and transcript components to approximately 15 in both microbiome and transcriptome.

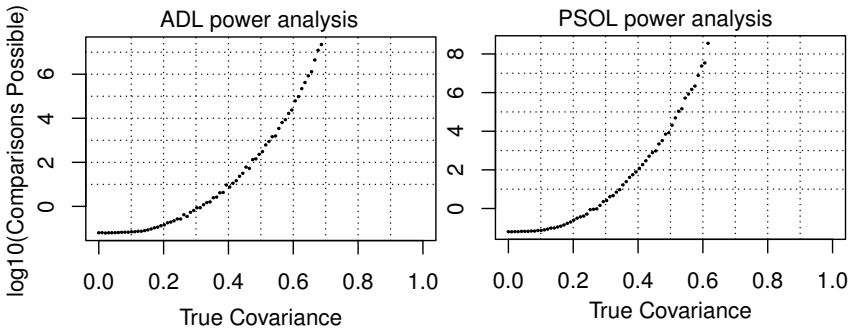

Figure 5.1: Power analysis. For an estimated true covariance of 0.5 and to retain 80% power, the approximated number of possible hypothesis tests after correcting for the FWER was estimated.

5.3.2 Integration pipeline

Given the limited number of pairwise tests available, a stringent dimensionality reduction scheme was applied to both the microbiome and transcriptome datasets. For both the microbiome and transcriptome, hierarchical clustering was applied to identify groups of similar features. Then, an appropriate representative from each cluster for linear modelling was selected (**Figure 5.2**).

First, dimensionality reduction of the transcriptome was performed in a semi-supervised manner. If microbiota are associated with inflammation and or diseased status, then it is likely that microbially induced transcriptional candidates will also be disease associated genes. The total number of transcripts on the array was initially reduced to those found to be differentially expressed between healthy and disease samples as described in (**Chapter 4**). This first reduction step reduced the total number of genes from 32633 to 1260 AD associated, and 2516 PSO associated transcripts. Next, DEGs were split into upregulated and downregulated groups and hierarchical clustering was performed using 1 - Pearson correlation as the distance measure with Ward's linkage method. The optimal cut height

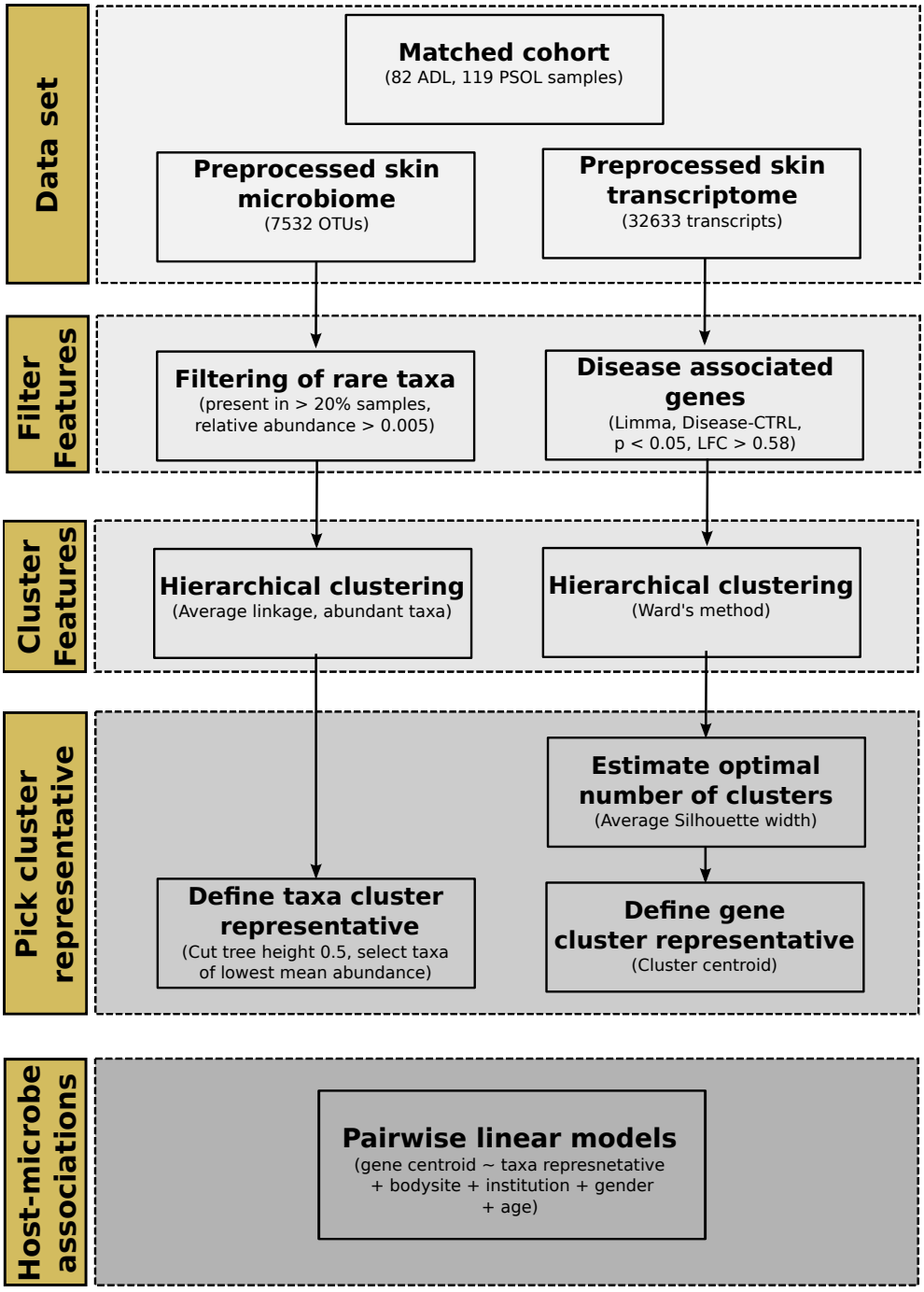

Figure 5.2: Flow diagram of integration pipeline.

was identified by clustering DEGs from $k = 2$ through 50 and picking k as the number of clusters which maximised the average silhouette width. For each cluster the centroid was calculated which was then used for testing microbe host-transcriptome associations in a linear model (**Figure 5.2**).

Next, the scale of unsupervised dimensionality reduction in the transcriptome was used to guide the magnitude of dimensionality reduction in the microbiome. OTUs were summed up at each taxonomic level and then filtered to retain only abundant taxa which were present in at least 20% of samples with a mean relative abundance of 0.005. Summation at each taxonomic level was performed to allow higher order taxa which may not be well represented at the OTU level under the stringent filtering criteria. Filtered taxa were then hierarchically clustering using 1 - Pearson correlation as the distance measure. The dendrogram was cut a height of 0.5 and the taxa with the lowest mean abundance was selected [29] (**Figure 5.2**).

5.3.3 Host-microbe associations in AD

Genes differentially expressed in lesional AD were identified with limma as described in (**Chapter 4**). According to the silhouette width, the optimal numbers was 9 clusters for upregulated genes (denoted as U-, silhouette width = 0.21), and 7 clusters for downregulated genes (labelled as D-, silhouette width = 0.17, **Figure 5.3 A-B)**. Abundant taxa were then hierarchically clustered and the dendogram was cut at a height of 0.5 revealing 15 clusters (**Figure 5.3 C)**. A cut height of 0.5 was selected as to keep the total number of hypothesis tests close to the approximate number of 230 possible pairwise tests estimated from the power analysis (**Section 5.3.1**). From each of these clusters, the representative with the lowest mean was selected, which for the majority of taxa was at the OTU level except for three at the genus level, one at the family level and one at the order level. Pairwise linear models were then implemented to determine the association between the 16 disease associated gene expression centroids and the 15 microbe cluster representatives.

Three significant associations were identified ($p < 0.05$, $r > 0.4$), all of which involved *Staphylococcus aureus* (**Figure 5.4 A)**. The strongest association identified was between *S. aureus* and the downregulated cluster D1, ($r = -0.49$, $p = 0.03$, **Figure 5.4 B)**. To identify top gene candidates within a cluster, the correlation of each member to the cluster centroid was calculated. Top genes associated with D1 (**Figure 5.4 E)**, included genes

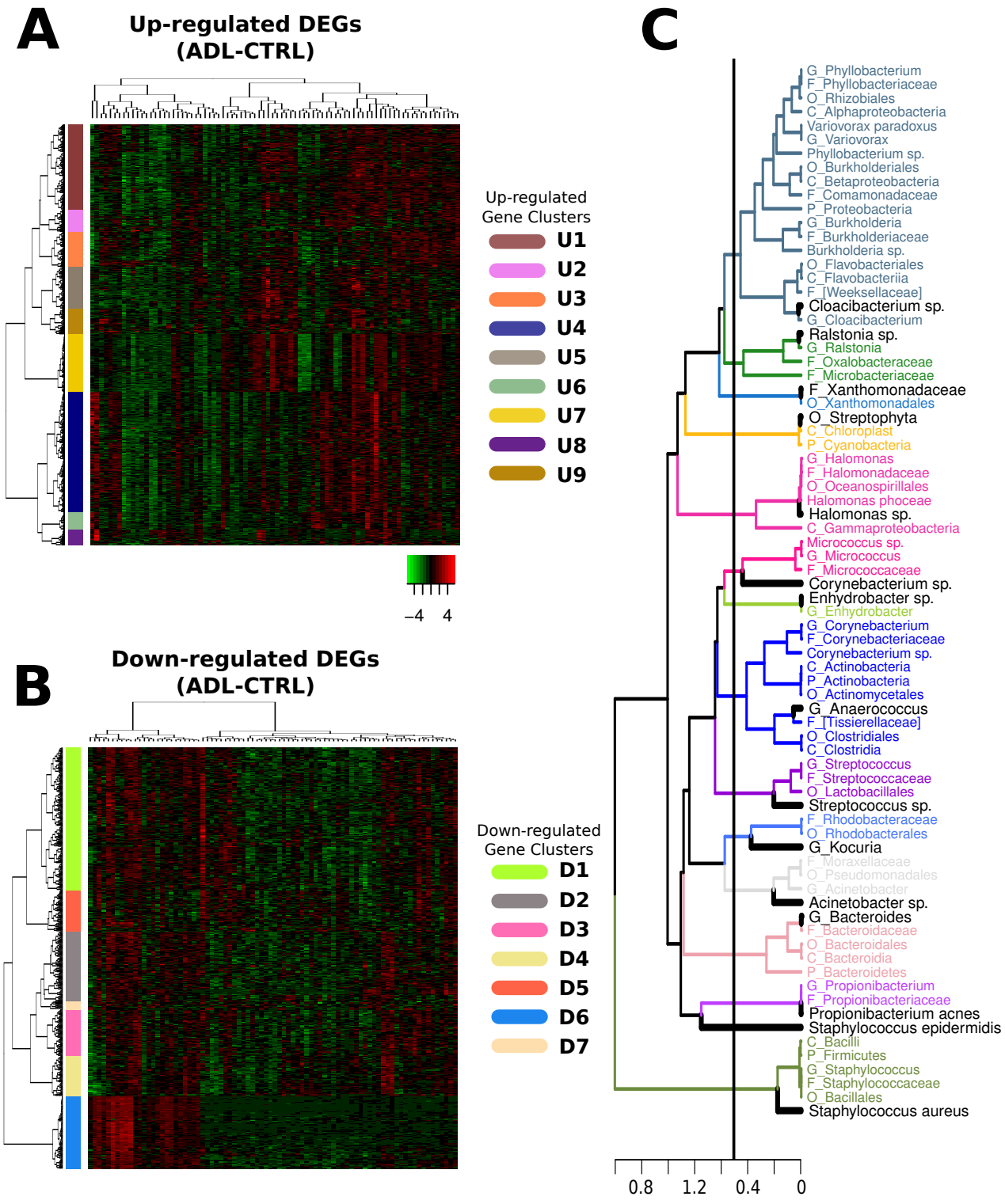

Figure 5.3: Dimensionality reduction of the Atopic transcriptome and microbiome. (A) The optimal number of clusters amongst upregulated genes is shown in a heatmap. Row colors correspond to clusters. (B) Clusters amongst downregulated genes. (C) Dendrogram of microbes. The black line corresponds to a cut height of 0.5. Font colour correspond to microbe clusters and the microbe in black for each cluster is the selected lowest mean representative.

corresponding to immunity such as IL34 and RORC which displayed some of the strongest (negative) correlations with *S. aureus*. IL34 is associated with the survival and differentiation of langerhans cells and has recently been implicated in lesional AD where it has been suggested that it may play a role in the inhibition of the inflammatory cascade [211]. The top enriched GO terms associated with D1 were 'circadian regulation of gene expression' (p = 2.0e-04) and 'rhythmic process' (p = 2.0e-04, **Table 5.2**). Several genes within D1 were central components of the mammalian circadian clock including PER1, PER3, CRY2 and CIART. Whilst these genes were not amongst that most highly correlated with the cluster centroid, they were amongst the group of genes which had the strongest correlation with *S. aureus*.

Staphyococcus aureus positively correlated with the U1 cluster (r = 0.46, p = 0.09, **Figure 5.4 C)**. The top enriched GO terms for U1 were 'positive regulation of NF-κB transcription factor activity' (p = 1.8e-05), 'positive regulation of defence response' (p = 3.4e-04), and 'regulation of cytokine production' (p = 7.6e-04, **Table 5.2**) demonstrating a heightened state of innate immune activity with increasing abundance of *S. aureus*. Several of the top genes within the U1 cluster encoded for antimicrobial peptides such as the members of the S100A family and DEFB4A, as well as other features of the epidermal compartment such as KRT6C and DSG3 (**Figure 5.4 F)**. The IL4R gene was also amongst the top *S. aureus* associated genes; the product of this gene has Th2 cytokines IL-4 and IL-13 as ligands [212].

The final *S. aureus* associated gene cluster was U9 (r = 0.44, p = 0.09, **Figure 5.4 D**). This cluster was enriched for immune system processes such as 'activation of immune response' (p = 1.9e-02, **Table 5.2**). Study into the composition of U9 revealed that immunological GO terms were enriched mostly due to the presence of five transcripts encoding components of the classical complement system pathway including C1QB, C1QA, CR1, C3AR1 and C1QC (**Figure 5.4 G)**. This was further supported by GO enrichment for 'complement activation' (p = 2.09e-02) which is known to play a role in opsonization of *S. aureus* [213]. Other top genes within U9 corresponded components of the ECM such as COL6A3, COL4A1 and TIMP1 (**Figure 5.4 E)** which also was indicated as an enriched GO term (**Table 5.2**).

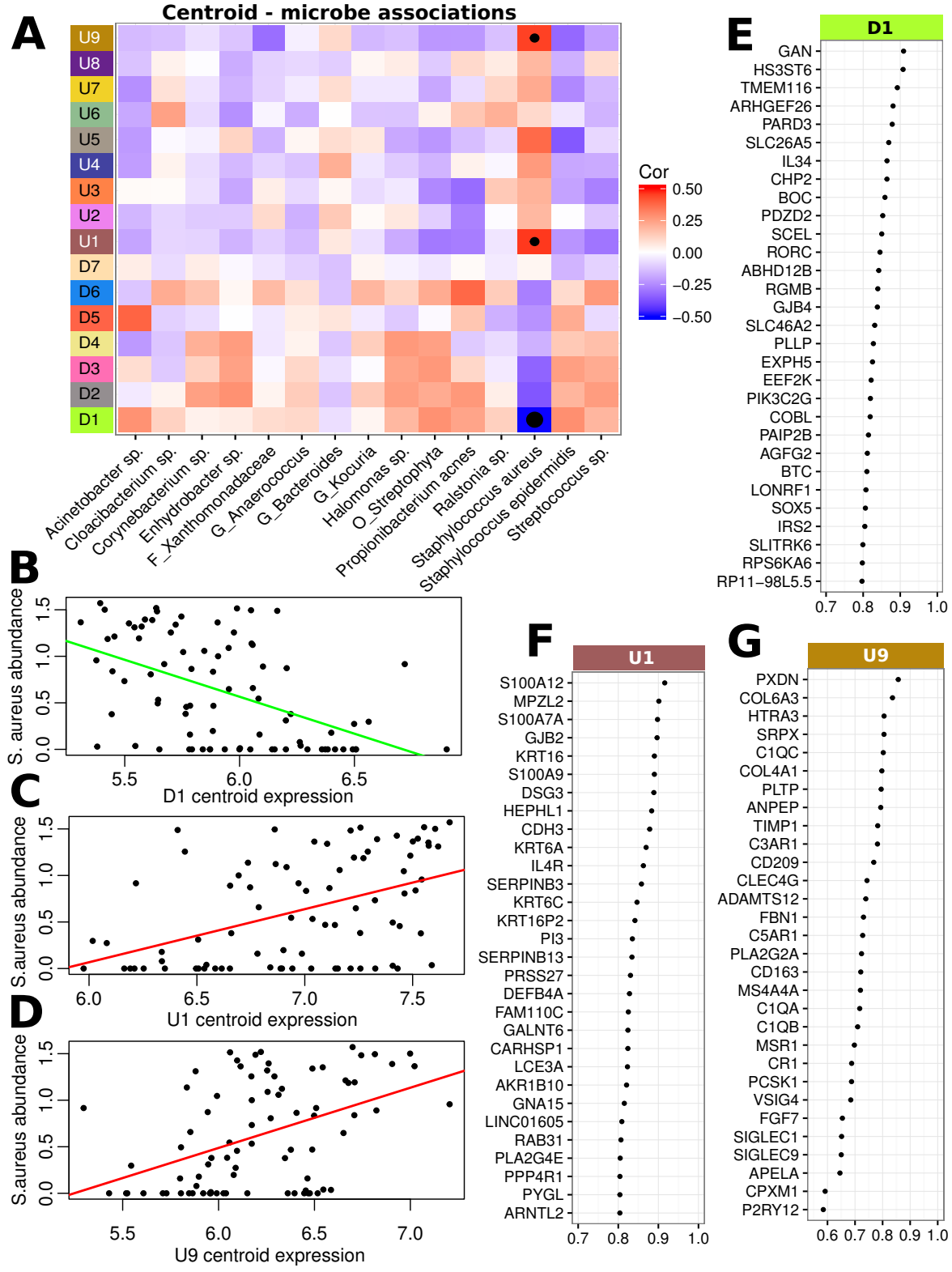

Figure 5.4: Associations between the ADL microbiome and transcriptome. (A) associations between gene clusters and microbe cluster representatives. Heatmap tile color corresponds to the correlation, and a point within a tile represents a significant association ($p < 0.1$, $r > 0.4$). (B) *S. aureus* vs D1 centroid expression. (C) *S. aureus* vs U1 centroid expression. (D) *S. aureus* vs U9 centroid expression. (E) The top 30 members of D1 measured by pearson correlation to the cluster centroid. (F) Top 30 genes for U1. (G) Top 30 for U9.

5.3.4 Host-microbe associations in PSO

The same exploratory analysis pipeline was applied in psoriasis to identify potential host microbe interactions. The initial 2516 DEGs optimally clustered into 6 upregulated (**Figure 5.5 B**), and 5 downregulated clusters (**Figure 5.5 C**). Next, taxa were hierarchically clustered and the dendrogram was cut at the same height of 0.5 used in AD which identified 25 microbe clusters (**Figure 5.5 A)**. As observed in AD, for most clusters, the lowest mean representative was selected at the OTU level resulting in 18 at the OTU level, 2 at the genus level and 5 at the family level. Pairwise linear models were then implemented to determine the relationship between the psoriasis associated gene expression centroids and selected taxa cluster representatives.

Cluster	Term	Adjusted.P.value
D1	circadian regulation of gene expression (GO:0032922)	2.0E-04
D1	rhythmic process (GO:0048511)	2.0E-04
D1	regulation of fat cell differentiation (GO:0045598)	2.0E-03
D1	regulation of phosphatase activity (GO:0010921)	3.9E-03
D1	regulation of dephosphorylation (GO:0035303)	9.9E-03
U1	positive regulation of NF-kappaB transcription factor activity (GO:0051092)	1.8E-05
U1	positive regulation of defense response (GO:0031349)	3.4E-04
U1	positive regulation of sequence-specific DNA binding transcription factor activity (GO:0051091)	7.9E-04
U1	regulation of cytokine production (GO:0001817)	7.9E-04
U1	positive regulation of cytokine production (GO:0001819)	9.7E-04
U9	activation of immune response (GO:0002253)	1.9E-02
U9	complement activation (GO:0006956)	2.9E-02
U9	protein activation cascade (GO:0072376)	3.2E-02
U9	extracellular matrix disassembly (GO:0022617)	3.8E-02
U9	regulation of extracellular matrix disassembly (GO:0010715)	4.3E-02

Table 5.2: Enriched GO terms amongst *S. aureus* associated gene clusters

A lack of concordance between the microbiota and host transcriptome in PSO was observed with no significant associations (**Figure 5.5 D)**. The correlations across all pairwise tests were low, and the maximum correlation was 0.26 between the family *Xanthomondaceae* and the upregulated gene cluster U6 (**Figure 5.5 E)**. In **Chapter 3**, analysis of the microbiota identified strong associations between *Corynebacterium simulans* and psoriasis status, however, no significant transcriptomic signal was identified for this species and the top correlation with any centroid was 0.14 (**Figure 5.5 F**). Overall, these results indicate a surprisingly low concordance between the psoriatic microbiota and transcriptome.

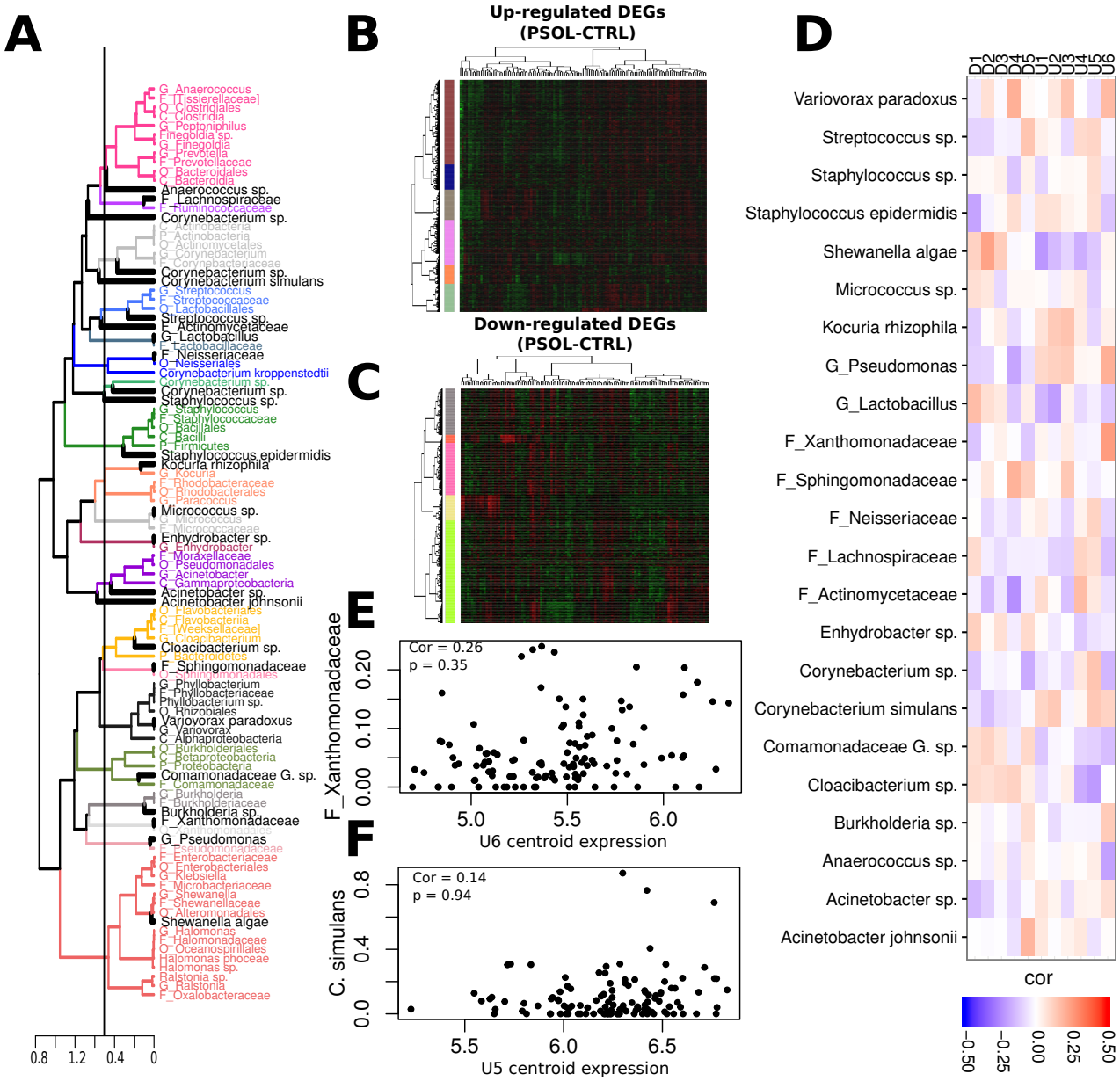

Figure 5.5: Dimensionality reduction and associations between the PSOL transcriptome and microbiome. (A) Dendrogram of microbes. The black line corresponds to a cut height of 0.5. Font colour correspond to microbe clusters and the microbe in black for each cluster is the selected lowest mean representative. (B) The optimal number of clusters amongst upregulated genes is shown in a heatmap. Row colors correspond to clusters. (C) Clusters amongst downregulated genes. (D) Associations between microbes and gene clusters. (E) *F. Xanthomonadaceae* vs U6 cluster expression. (F) *C. simulans* vs U5 cluster expression.

5.3.5 Covariation of *Staphylococcus aureus* with transcriptome factors

To further analyse *S. aureus* associated transcriptomic signatures, an unsupervised feature reduction of the transcriptome was performed using principal component analysis. The complete AD transcriptome (32633 transcripts) was reduced to genes with greater than the median variance leaving 16316 genes. Principal component analysis was performed to further reduce 16316 variable genes to 17 factors which explain 50% of the transcripomic variability in AD [29]. Linear models were then utilised to determine the association between principal components and *S. aureus* relative abundance.

One significant association was identified with PC1 (**Figure 5.6 A)** which explained 9.5% of the variation in ADL transcriptome. *S. aureus* was positively correlated with PC1 (cor = 0.42, p = 0.07, **Figure 5.6 C)** indicating that the transcriptomic architecture underlying AD and *S. aureus* are closely related. To investigate the functional properties of PC1, the top 20 positive and negative loadings were analysed (**Figure 5.6 B**). The top loadings correspond to genes which are the most highly weighted and represent the most variable genes along the direction represented by the principal component. Within the top loadings of PC1, genes corresponding to the epidermal compartment such as DSG3, DSC2, KRT6A, KRT16 and KRT6C were positively weighted. The Th2 cytokine receptor IL4R was also found to the amongst the top weighted genes. In the negative direction, IL34 and RORC were amongst the most weighted, further supporting a negative association between *S. aureus* and these genes.

To assign functional annotations to PC1, a pre-ranked gene set enrichment analysis [147] of the REACTOME database [145] was performed using the PC1 loadings (**Figure 5.6 D)**. As expected, this analysis revealed a strong enrichment for immune categories such as 'Immune System' and 'Adaptive Immune System' in the positive direction. In the negative direction, 'Generic transcription pathway' as well as 'Metabolism of Lipids and Lipoproteins' were enriched. These results demonstrate that *Staphylococcus aureus* is positively associated with heightened immune system activity and may be associated with deficiencies in lipid metabolism, a process known to be characteristic of AD [58]. Many of the top weighted genes in PC1 were also found to be significantly associated in the exploratory analysis described in (**Section 5.3.3**) further supporting a host-pathogen interaction.

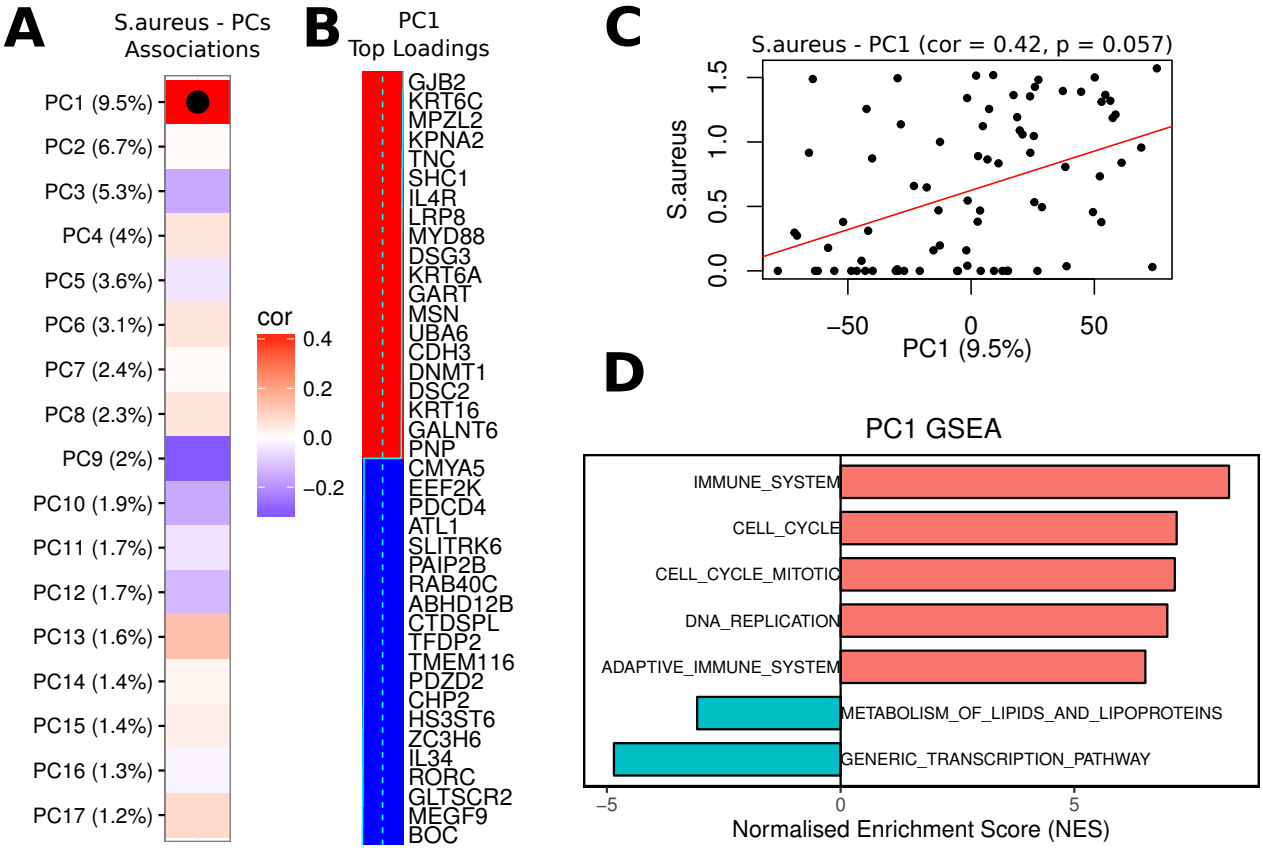

Figure 5.6: Association between principal components and *S. aureus*. (A) PC1 significantly correlated with *S. aureus* relative abundance ($p < 0.1$, $r > 0.4$) Heatmap tile colour corresponds to the correlation and a point within a tile indicates a significant association. (B) Top 20 positive and negative loadings of PC1. (C) *S. aureus* abundance vs PC1 scores. (D) Pre-ranked GSEA of REACTOME pathways using PC1 loadings.

5.3.6 A *Staphylococcus aureus* transcriptomic signature

In order to determine a direct *S. aureus* transcriptomic signature. The total 82 ADL samples were dichotomised into equal groups of 41 samples based upon the median abundance of *S. aureus* (**Figure 5.7 A)**. Differential analysis of the two groups, '*S. aureus*-High' and '*S. aureus*-Low', was performed with limma [137] which revealed a signature of 578 genes (330 upregulated, 248 downregulated, $p < 0.1$, **Figure 5.7 B**). The top upregulated genes included the antimicrobial peptides S100A7A, S100A9, S100A8, S100A12 and DEFB4A as well as the Th2 cytokine receptor IL4R, and components of the ECM including COL6A3, COL4A1 and COL4A2, many of which were also identified in the exploratory analysis. Genes downregulated in the presence of *S. aureus* included those associated with the circadian clock such as RORC, PER and CRY2 as well as the cytokine IL-34.

Enrichment of upregulated genes with Ingenuity Pathway Analysis [143] identified the top pathway as 'Hepatic Fibrosis/Hepatic Stellate cell activation' ($p = 0.015$, **Figure 5.7 C**). This pathway included genes encoding for collagen, COL4A1, COL4A2 and COL4A2, as well as profibrotic immune genes TGFB1, IL4R and MMP1 suggesting that *S. aureus* may be associated with ECM remodelling in the dermis which is thought to play a role in inflammation [214]. The 'role of IL17A in psoriasis' pathway was also enriched, however, this pathway was enriched due to up-regulation of the S100A8, S100A9, S100A7A and DEFB4A antimicrobial peptides and not due to differential expression of IL17A. Enrichment of downregulated genes revealed only weak associations. 'PXR/RXR activation' was enriched ($p < 0.1$, **Figure 5.7 D**) and may reflect an association between *S. aureus*, skin dryness and epidermal lipid deficiencies [59, 58]. As a separate isolated test, the stratified '*S. aureus* high' group cohort was shown to have higher expression levels of PC1. Lastly, differences in disease severity was tested between the '*S. aureus* high' and '*S. aureus* low' groups. Patients with high abundances of the '*S. aureus*-high' group had a significantly higher Local SCORAD than those in the '*S. aureus*-low' group (**Figure 5.7 E**, $p = 0.006$).

This stratification analysis was also performed for PSO candidates such as *C. simulans* and *C. kroppenstedtii* in PSO lesional samples which revealed no significant genes.

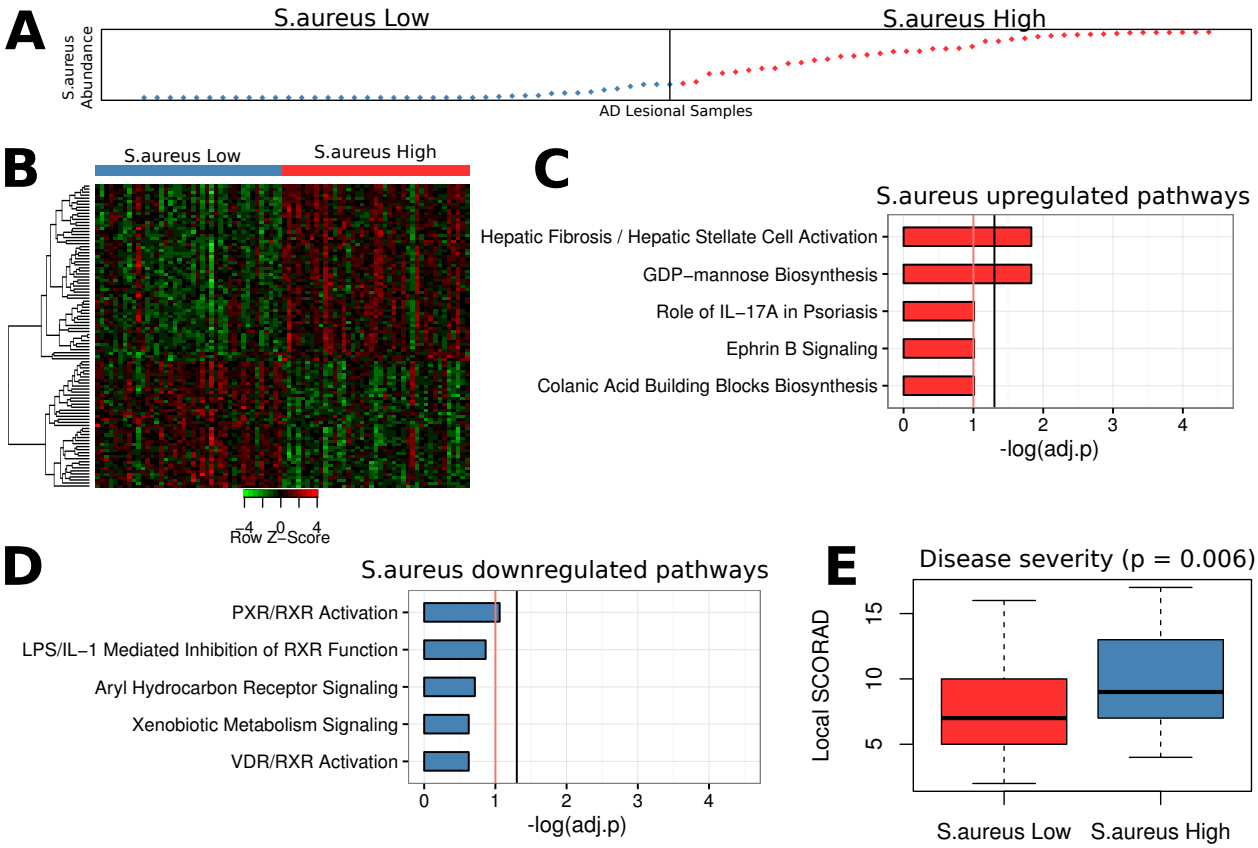

Figure 5.7: Transcriptome stratification analysis. (A) Samples were stratified into *S. aureus*-high and *S. aureus*-low groups based upon median abundance. (B) Heatmap of differentially expressed genes between *S. aureus*-high and *S. aureus*-low. (C) Upregulated pathways as identified by IPA. The black line corresponds to p = 0.05, and the red line corresponds to p = 0.1. (D) Downregulated pathways. (E) Association with Local SCORAD tested with Wilcoxon ranked sum test.

5.4 Conclusions and Discussion

Several studies have analysed both the skin inflammation associated host transcriptome and microbiome, however, this is first time that both datasets have been simultaneously analysed and integrated to identify potential host-microbe interactions. It has been hypothesised that psoriasis may be triggered by microbiota stimuli due to its overall similarity to Crohn's disease [215], in which chronic inflammation is thought to be associated with dysbiosis in the intestine [80]. Mutations within genes of both the innate and adaptive immune system, as well as the epidermal barrier are characteristic of PSO [73] therefore, it was surprising to find a lack of concordance between taxa on psoriatic skin and the expression of disease associated transcripts. In **Chapter 3** the psoriatic microbiota was unchanged between uninvolved and lesional states, yet in **Chapter 4**, extreme transcriptional differences were identified. The lack of concordance observed in the current chapter may suggest an inability of the host transcriptome to shape community composition (or vice versa) in the short term, and it may be the case that changes are accumulated over the course of disease development.

Staphylococcus aureus is a known pathogen infecting the lesions of patients with AD as confirmed in **Chapter 3**. Here using several integrative approaches, it can be seen that the relative abundance of *S. aureus* is clearly related to significant proportions of the atopic transcriptomic signature. *S. aureus* positively correlated with proinflammatory gene clusters which define AD, and negatively correlated with genes downregulated in lesions. Perhaps the most interesting finding, was the positive association with IL4R which is the receptor of the Th2 cytokines IL4 and IL13 [212] which are of key importance in the pathogenesis of AD [44]. It was also clear that *S. aureus* was associated with the innate immune system via expression of anti-microbial peptides as well as the complement system. Whilst host defence mechanisms were of heightened activity, patients with greater abundance of *S. aureus* were also characteristic of severe disease. *S. aureus* was also positively associated with components of the extracellular matrix such as COL6A3, COL4A1 and TIMP1. The ECM plays important roles involving immune cell activation, proliferation and migration [214] and it may be that *S. aureus* possesses pro-fibrotic properties, or is associated with Th2 activity which can induce fibrosis and is associated with AD [208].

Interestingly, the strongest association found was an inverse correlation between genes in a cluster enriched for circadian clock genes. Circadian rhythms have recently been shown

to control the expression of proinflammatory cytokines and that mutations within clock genes results in a psoriasiform inflammation in mice [191]. Furthermore, IL34 was inversely correlated with *S. aureus* which is important for the development of Langerhans cells [216]. Overall, whilst no significant associations were detected in psoriasis, the results presented in this chapter provide an insight into the transcriptional imprint of *Staphylococcus aureus* in AD and provides a foundation for further mechanistic studies.

Chapter 6

Co-expression networks analysis of skin inflammation

6.1 Introduction

Understanding transcriptional dysregulation is key to unravelling the mechanisms which cause disease. Differential analysis provides only a reductionist view of the information contained within large scale transcriptome datasets. In such analysis, genes are considered as independent entities whereas in reality, genes together with other biomolecules interact within highly complex and intricate systems.

One method to investigate biological systems is to model gene expression data as a graph in which nodes correspond to transcripts, and edges correspond to the interactions between gene pairs allowing the relationships between genes to be studied. Interaction networks facilitate the analysis of systems using techniques based upon graph theory, such as clustering or centrality measures for hub gene detection. The concept of cluster detection, also known as community or module detection can be used to identify sets of genes which may be involved in similar biological processes. Thus, it is a powerful technique to identify new candidate genes by exploiting the principle of guilt by association [217, 218].

The goal of transcriptional network reconstruction is to infer gene-gene interactions directly from gene expression measurements. Several methods for network inference have been proposed in literature. The simplest approaches calculate a similarity measure between genes, typically via correlation, where an edge connects gene pairs if the correlation satisfies a

certain threshold [219]. Other methods to reconstruct gene-gene interaction networks include those which calculate mutual information between gene pairs such as ARACNe [220], and others such as those based upon partial correlations [221] or bayesian methods [219]. Despite several available approaches, the high dimensionality of transcriptome data has resulted in popularity for correlation based methods as they are intuitive and computationally inexpensive compared to partial correlation and Bayesian methods which do not scale well with large datasets [222]. The most established co-expression network inference method, Weighted Gene Co-expression Network Analysis (WGCNA), has been used to reconstruct co-expression networks numerous times since its initial publication in 2005 [150]. WGCNA has been applied to transcriptomics data in many contexts. These include, relating co-expression modules to specific brain regions [223], between species [224], in diseases such as Huntington's disease [225], autism [226], heart disease [227], hepatitis [228] and cancer [229, 230]. Furthermore, in a recent study, the WGCNA and ARACNe methods were suggested to be the most robust methods for constructing global co-expression networks [222].

The analysis presented in this chapter investigates the reconstruction of skin associated co-expression networks within the WGCNA framework [150, 231]. The reconstructed networks were then used to identify genes and gene modules which play important roles in skin inflammation. By performing a large scale genome-wide integrative network analysis, the objective was to identify interacting communities of genes to find processes which may be relevant in the pathogenesis of Psoriasis (PSO) or Atopic dermatitis (AD). Identifying gene communities significantly reduces dimensionality allowing host transcriptomic profiles to be integrated with microbial abundance to find modules which covary with pathogens such as *Staphylococcus aureus* or *Corynebacterium simulans*, as well as those which are associated with clinical severity.

Independent co-expression networks were reconstructed for each cohort, and then two main analyses were performed. In the first, module preservation analysis [232] was performed to identify differentially co-expressed modules between healthy and diseased states. The second involved associating co-expression modules to clinical variables and a selection of potentially pathogenic microbes in order to uncover host-microbe interplay.

6.2 Methods

6.2.1 Data selection and preprocessing

Quality controlled and RMA normalised gene expression data was obtained from Insititut Curie as well as quality controlled and normalised 16S microbiome sequencing data was obtained from the Karolinska institutet and Institut Curie as part of the MAARS consortium project. Details of patient recruitment, sampling and data processing are described in **(Sections 2.3.1, 2.3.2 and 2.3.3)**. As a major objective of this analysis was to identify host-microbe interactions, only samples with both microbiome and transcriptome samples were considered as described in **Table 6.1**.

The objective of comparative network analysis was to compare the structure of gene-gene correlations between lesional, non-lesional and control networks, therefore, for each disease (AD, PSO) a body site matched cohort was defined to remove under-represented body sites (in either the control or disease groups). This step was performed to reduce potential sources of non-clinically associated variation, so that when disease and control groups are compared, biopsies from specific body sites in each network are more appropriately balanced. For Atopic dermatitis, ADL, ADNL and an AD body site matched control group (CAD) were defined from upper back and thigh samples. The psoriasis datasets, PSOL and PSONL as well as a body site matched PSO group (CPSO) were defined from lower back and upper back samples.

6.2.1.1 Transcriptome preprocessing

Co-expression analysis can be sensitive to array outliers, therefore, the initial step in the WGCNA pipeline is to remove potential array outliers. For each cohort (ADL, ADNL, CAD, PSOL, PSONL, CPSO), the Euclidean distance between samples was calculated and samples were clustered with average linkage hierarchical clustering. Samples which did not cluster within the main body of samples (**Appendix Figure B.1 - B.2**) were removed. A complete overview of the samples used in network construction are described in **Table 6.1**. Genes with less than the median variance across all cohorts were removed ensuring that networks for each cohort were constructed on the same set of genes.

6.2.1.2 Microbiome preprocessing

For matched microbiome samples, a variance stabilising arcsine square root transformation [5] was applied to OTU relative abundances. Specific microbes of relevance in the inflammatory skin microbiome were selected and tested for association with the transcriptome including *Corynebacterium simulans*, *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Propionibacterium acnes*, *Peptostreptococcus anaerobius*, and *Neisseriaceae. G. sp.*

6.2.2 Network construction and module detection

6.2.2.1 Definition of the adjacency matrix

Networks were reconstructed independently for each cohort within the WGCNA framework [150, 231]. First a similarity measure was calculated between all gene pairs resulting in a correlation matrix between all genes. Typical correlation measures include Pearson, Spearman or the biweight midcorrelation (bicor). The biweight midcorrelation is based upon the sample median instead of the mean and is less sensitive to outliers [233]. A signed co-expression network which retains information regarding the direction of co-expression was constructed, thus the correlation between nodes i and j was defined as:

$$c_{ij} = |0.5 + 0.5 \times bicor(x_i, x_j)| \tag{6.1}$$

where x_i and x_j are the expression profiles of nodes i and j across all samples.

Setting the threshold for the correlation matrix using a hard threshold, for example, connecting gene pairs with correlation strengths greater than 0.8 may result in loss of information as a gene pair with correlation of 0.79 would not be connected [150]. To overcome this, a weighted adjacency matrix is defined [150]. The weighted adjacency defines the connection strengths between genes. For each network, the signed correlation coefficients were iteratively raised to a power β through values 2 to 30. The soft thresholding parameter, β, is applied because transcriptome data is noisy and correlations may arise due to technical error [231]. Higher values of β exponentially penalise weaker correlations by shrinking them closer to 0 and emphasises stronger correlations by influencing them less. Biological networks are assumed to be of scale free topology where the degree distribution of a network follows a power law [234, 218]. To test for a scale free topology, the degree distribution (fraction of nodes which have degree k) is plotted against k in a log-log plot.

If a straight line with a negative slope is observed, the network topology approximates a scale free topology, i.e., many nodes with low connectivities and characterised by few highly connected hubs. The parameter β is selected to ensure that network adopts a scale free topology. Hence, the adjacency of nodes i and j is defined as:

$$a_{ij} = c_{ij}^{\beta} \tag{6.2}$$

Next the Topological Overlap Measure (TOM) [150] is defined. As the overall objective of WGCNA is to find modules of genes which are highly interconnected, i.e., genes which share many of the same interaction partners, a node similarity measure is calculated which quantifies how similar two nodes a, b are with respect to how similar the neighbours of a are to the neighbours of b. TOM takes into account the connection strengths between gene pairs whilst also considering connection strengths between paired neighbours. Topological overlap also has the added benefit of reducing the influence of spurious correlations due to random noise. TOM is defined as:

$$TOM_{ij} = \frac{|a_{ij} + \sum_{u \neq i,j} a_{iu} a_{uj}|}{min(k_i, k_j) + 1 - |a_{ij}|} \tag{6.3}$$

where k_i and k_j are defined as the total connectivities of nodes i and j:

$$k_i = \sum_{u \neq i} |a_{ui}| \tag{6.4}$$

For a gene pair, TOM values lie between 0 (completely unconnected) and 1 (fully connected) and the TOM dissimilarity matrix (1-TOM) is used in conjunction with a community detection algorithm to identify groups of highly interconnected genes. Average linkage hierarchical clustering was performed and branches of the dendrogram were clustered into co-expression modules with the cutreeDynamic function [235]. The dynamic tree cut method does not cut at a static threshold and independent branch cutting is performed based upon branch shape as a static tree cut cannot identify nested modules of genes [235].

6.2.2.2 Module eigengenes

For each module after branch cutting, the module eigengene (ME) [236, 237] is calculated. The module eigengene is the first right singular vector of the standardised module expression matrix [238] and is equivalent to the first principal component. Module detection can result in multiple modules which express similar expression patterns and the authors

suggest merging modules to reduce redundancy [238]. The pairwise Pearson correlation between module eigengenes was calculated and modules with similar expression patterns corresponding to a correlation of greater than 0.8 were merged. After merging, module eigengenes were recalculated using the merged module definitions. To ensure modules constituted only genes with similar expression patterns, all genes were correlated to the module eigengene, and those with a correlation (kME) < 0.4 were reassigned to the 'gray' module for genes with no clear module membership (see Section 6.2.2.4). For readability reasons only, modules in every network were relabelled such that they matched the colour of the module with the most significant overlap defined by a Fisher's exact test within the ADL network.

6.2.2.3 Functional enrichment

Modules were annotated using functional enrichment analysis. Unless otherwise stated, gene ontology functional enrichment of biological processes was performed with clusterProfiler [239]. Further analysis was performed with IPA [143] and consensusPathDB [240] to obtain knowledge of enriched pathways and upstream regulators. P values were adjusted using the Benjamini Hochberg method.

6.2.2.4 Identification of hub genes

Hub genes were defined using the eigengene based connectivity measure kME [241], also known as module membership. kME is calculated as the correlation of the expression profile of gene i with a module eigengene ME and thus represents the extent to which a gene follows the expression profile of a module. Genes with a high kME, i.e., close to 1 tend to be highly connected to genes within a module and have a similar expression profile to the module eigengene. To determine the importance of a gene within a module, genes were ranked on their kME values.

Alternatively, hub genes can be identified using connectivity based measures. The whole network gene connectivity [150] or degree k of a node i is calculated as:

$$k_i = \sum_{j \neq i} a_{ij} \tag{6.5}$$

Module hub genes, can be identified by Intramodular connectivity, kIM [150]. Intramodular connectivity is defined as the sum of the weighted adjacencies between gene i in module

m and all other genes within module m.

$$kIM_i = \sum_{j \in m} a_{ij} \tag{6.6}$$

To enable comparison of kIM across modules, kIM was scaled by the gene with the maximum connectivity. Reports show that kIM and kME are highly related [231].

6.2.2.5 Differential connectivity

Differential connectivity of genes between networks was performed to identify differentially regulated hubs using the method described in [151]. Within each network the connectivity of each node was calculated as the sum of all weighted adjacencies with other genes in the network. Then, the the scaled weighted connectivity of node i was calculated as the connectivity of node i divided by the connectivity of the most highly connected node in the network. The difference in weighted scaled connectivities, kDiff, between networks A and B was calculated as kA - kB. Using the thresholds suggested by [151], differentially connected genes were considered those for which kDiff > 0.4.

6.2.3 Statistical analysis

To determine if modules overlapped with differentially expressed genes, differential analysis was performed using limma as described in (**Chapter 4**). Association between module eigengenes and traits was performed using a linear model to account for sources of extraneous variation. The lm function in R was used with the formula, ME $\sim$ trait + anatomical location + institution + age + gender. The traits evaluated included clinical severity (SCORAD, PASI) and arcsine square root transformed OTU relative abundances [5]. Significant associations were those with a Benjamini Hochberg adjusted p value < 0.1 with a correlation > 0.3. Correlation coefficients between traits and eigengenes were calculated with Pearson correlation.

6.2.4 Network visualisation

Network subgraphs for module preservation analysis were created by selecting edges with correlation > 0.65 between the top 25 hub genes as identified by kME. Module subnetworks were visualised in either Cytoscape [168] or Igraph [242]. For the network diagrams of *S.*

aureus associated modules, the top 20 hub genes identified in each module by kME were selected and the top 10% of edges were retained.

6.2.5 Module preservation

Module preservation analysis was performed using two main categories of preservation statistics. The first, and most naive approach, is the application of cross tabulation statistics, where given module definitions in two networks, the overlap of genes across modules was evaluated with Fisher's exact tests. This approach is intuitive for observing a general overview of module preservation between two networks, however, cross tabulation statistics do not take into account information regarding the correlations, or density of connections between modules [232]. Furthermore, as cross tabulation statistics require module definitions in both networks, these are likely to be somewhat dependent on the network construction procedure and parameter selection in which the modules were defined [232].

The second approach corresponds to network based statistics. Network based preservation statistics require module definitions in only one network, known as the reference network. To assess module preservation, the module definitions and adjacency matrix for the reference network are used to determine if clusters can be identified within the adjacency matrix of the test network. This addresses the pitfall that preserved clusterings can only be identified based upon the parameters of the module detection algorithm [232]. Two types of network based preservation statistics were considered: 'density' and 'connectivity' based. Density preservation determines if the connections within a module are of similar structure and quantity, whereas connectivity statistics are used to test if the interaction patterns between nodes are retained across two networks. The network based statistics used in this analysis are summarised in the following sections, however, for a detailed description refer to [232].

6.2.5.1 Density based preservation metrics

Density measures assume that if the density of connections within the test network are high, then it is likely that the reference module is preserved within the test network. The following density measures were considered [232]:

meanCor measures the correlation density, and is defined as the mean correlation of all

genes within the test network module, multiplied by the sign of the correlation in the reference network module. When the sign of the reference and test network are opposing, this decreases the value for meanCor.
meanKME is calculated in a similar way to meanCor however the correlation between neighbours is replaced with the correlation to the module eigengene, kME.
meanAdj is calculated as the mean adjacency of a module in the test network, thus, meanAdj represents the adjacency density. A module with a high mean adjacency in the test network suggests that the module is preserved.
propVarExpl uses the proportion of variance explained by the module eigengene as a density measure. The kME values are derived for the genes in a module by correlation with the module eigengene. The propVarExpl measure is defined as the mean squared kME values within the test network. [232]

6.2.5.2 Connectivity based preservation metrics

Connectivity measures work on the principle that the connectivity profiles of nodes within a reference network module should be similar within a test network if a module is preserved. To estimate the concordance, the correlation of connectivity measures between reference and test networks is calculated, thus, a high correlation indicates a preservation of node specific properties. The following connectivity preservation measures were considered [232]:

cor.cor corresponds to the correlation preservation which measures the concordance of correlations between genes in a co-expression network module. For genes in a module, the correlation matrix is defined in both reference and test sets. Cor.cor represents the correlation between the vectorised reference and test correlation matrices. For a highly preserved module, the correlation coefficients between genes should remain similar across reference and test networks.
cor.kIM refers to intramodular connectivity preservation. cor.kIM calculates the correlation between intramodular connectivities, defined in **equation 6.6**, in the test and reference networks. For a preserved module, hub genes within the reference network should remain highly connected within the test network, thus, one would expect a high correlation between intramodular connectivities.
cor.kME refers to the correlation of module memberships. The correlation of each gene and the module eigengene kME is calculated in both reference and test networks and the correlation is computed. Highly connected genes tend to have high kME values,

whereas peripheral genes tend to have low or negative kME values thus one would expect a preserved module to have a high cor.kME. [232]

6.2.5.3 Significance of module preservation

To assess the significance of module preservation, the observed value, *obs*, of a preservation metric a, is calculated between the reference and test network. Next the module definitions within the test network are randomly permuted n_{perm} times to construct an empirical null distribution of random modules of the same size. Then for the network statistic a, the mean and standard deviation of the permuted values are used to define a Z score [232]:

$$Z_a = \frac{obs_a - \mu_a}{\sigma_a} \tag{6.7}$$

To capture the degree of preservation represented by multiple connectivity and density statistics, the following composite measures are calculated.

$$Z_{density} = \text{median}(Z_{meanCor}, Z_{meanAdj}, Z_{propVarExpl}, Z_{meanKME}) \tag{6.8}$$

$$Z_{connectivity} = \text{median}(Z_{cor.kIM}, Z_{cor.kME}, Z_{cor.cor}) \tag{6.9}$$

and finally for an overall general view of module preservation, the $Z_{summary}$ score is calculated as the average of $Z_{density}$ and $Z_{connectivity}$. The authors of [232] suggest that a Z score > 10 indicates strong preservation, < 10 indicating weak preservation and < 2 that the module is not preserved. As there is more power to detect connectivity patterns between large numbers of genes compared to smaller numbers [232], the Z score is highly associated with module size. In the current analysis where modules vary in size, it can be more appropriate to measure preservation using the ranks of the obtained values. The *medianRank* was used to supplement $Z_{summary}$ statistics which are calculated in the same way as described above but with observed statistic rankings instead of permutation Z scores.

Table 6.1: Matched body site and network construction cohorts

		ADL	ADNL	CAD	CPSO	PSONL	PSOL
Patients (n)		74	72	107	101	96	98
Samples (n)		74	72	107	101	96	98
Gender (n)	Female	33	35	68	66	19	22
	Male	41	37	39	35	77	76
Anatomical location (n)	Lower back	0	0	0	89	76	89
	Thigh	41	38	95	0	0	0
	Upper back	33	34	12	12	20	9
Institution (n)	HHU	30	27	31	30	43	41
	KINGS	13	12	44	38	37	40
	UH	31	33	32	33	16	17
Age	Mean	43.4	43.4	34.3	35.3	47.1	47.8
	SD	13.3	14.3	12.7	13.8	13	13.5

6.3 Results

6.3.1 Data selection and preprocessing

Six individual networks were constructed corresponding to ADL, ADNL, Control-AD (CAD), PSOL, PSONL and Control-PSO (CPSO). Network construction was limited to arrays with available microbiome samples and AD networks were restricted to thigh and upper back, and PSO networks were constructed on lower and upper back samples. Next, to ensure co-expression networks were robust, samples which did not cluster within the main body of the samples were identified and removed by average linkage hierarchical clustering of the Euclidean distance matrix. The dendrograms and sample removals are shown in (**Appendix B, Figures B.1, B.2**). After sample removal, a total of 74 ADL, 72 ADNL, 107 CAD, 101 CPSO, 96 PSONL and 98 PSOL samples were used for network construction. A complete description of samples is shown in (**Table 6.1**) Next, the transcriptome was filtered to remove genes with low variability across all cohorts. All of the 32,632 genes present on the array are unlikely to be expressed in the skin, and genes that are of low variability are unlikely to correspond to systematic differences between diseases. Furthermore, construction of a genome wide co-expression network is computationally expensive, therefore, network construction was restricted to the top 50% of variable genes leaving a total of 16316 genes.

Table 6.2: Global network statistics

Network	SFTfit	Beta	Slope	Modules	Med connectivity
ADL	0.852	12	-2.12	22	15.8
ADNL	0.900	12	-2.72	13	15.5
CAD	0.943	12	-2.58	10	12.2
CPSO	0.990	12	-2.17	10	12.0
PSONL	0.980	12	-2.24	11	10.9
PSOL	0.902	12	-3.08	16	11.3

6.3.2 Network construction and module detection

Weighted gene co-expression networks require the selection of the correlation matrix exponent β. To identify the optimal value for the soft power threshold β, the signed biweight midcorrelation (bicor) matrix was calculated for each network and was then iteratively raised to powers 3 through 30. For each value of β, the degree distribution $p(k)$ and connectivity k of the network is calculated and a linear regression is performed to quantify the fit of the network to the scale free topology criterion, i.e., following a power law with a negative slope. To allow for comparisons between networks, a common value of β was selected for which all networks were approximately scale free with comparable median connectivities. A β value of 12 was selected as at this value the connectivity of all networks fit the scale free topology criterion with an R^2 of greater than 0.85 (**Figure 6.1**). Scale free fit, median connectivity and general network statistics are presented in (**Table 6.2**).

The topological overlap matrix was calculated to highlight gene pairs with similar interacting partners, and modules were defined using the cutreeDynamic method [150] independently in each network. Within each subnetwork, the module eigengene was calculated and pairwise correlations between module eigengenes was performed to identify highly related modules. As modules with similar expression patterns may represent redundancy, subnetworks with a correlation of greater then 0.8 were merged into a single module. Modules were labelled with a colour for interpretation, and genes which did not belong in any module were grouped into an unassigned 'grey' module. Modules for each network were then matched to modules in the ADL network by calculating the overlap in gene members and performing Fisher's exact tests. Modules with the most significant overlap were relabelled to have the same colour. As labels are arbitrary, this step has no significance on the analysis, and only aids the interpretation of modules across networks.

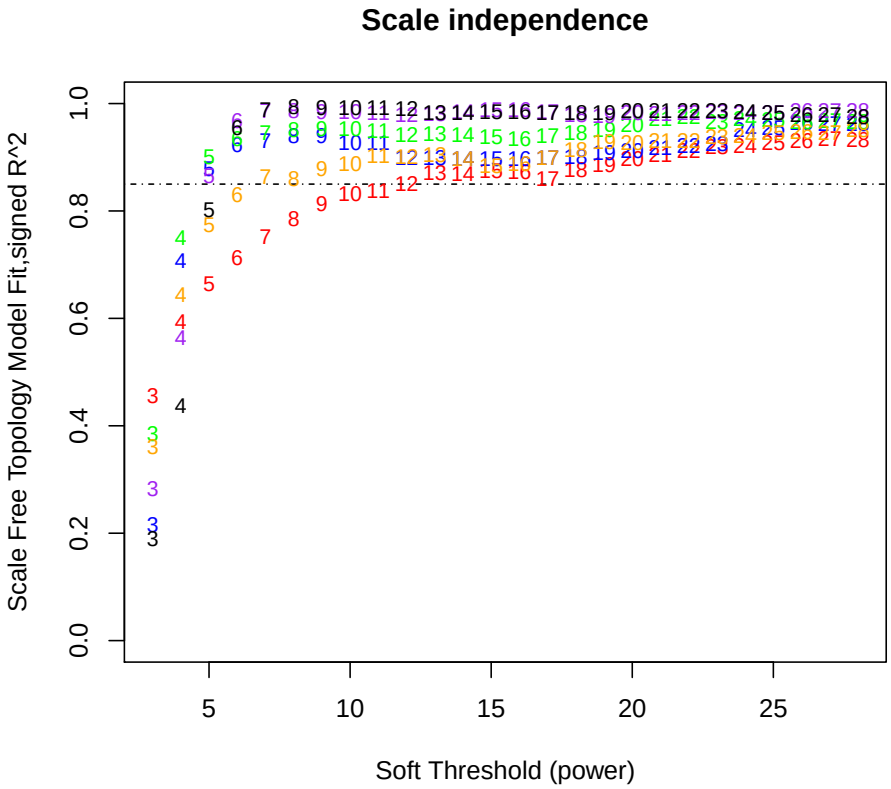

Figure 6.1: Scale free topology fit for all cohorts at varying levels of β. The horizontal line corresponds to a scale free topology fit of 0.85. Colours correspond to the cohort and at β = 12, all cohorts fit the scale free topology criterion > 0.85.

6.3.3 Atopic Dermatitis associated networks

Construction of the ADL, ADNL and CAD networks revealed 22, 13 and 10 modules respectively (**Figure 6.2 A**). The final networks and module definitions showed that networks were comparable in terms of slope and median connectivity (**Table 6.2**). Qualitatively, a concordance in module assignments can be observed across the ADL, ADNL and CAD networks by the tracks below the gene dendrogram (**Figure 6.2 A**).

6.3.3.1 CAD network modules

Modules identified within the skin co-expression networks must be annotated with a biological function to aid in interpretation. An enrichment of GO biological processes in each module using clusterProfiler [239] was performed which revealed a set of modules enriched

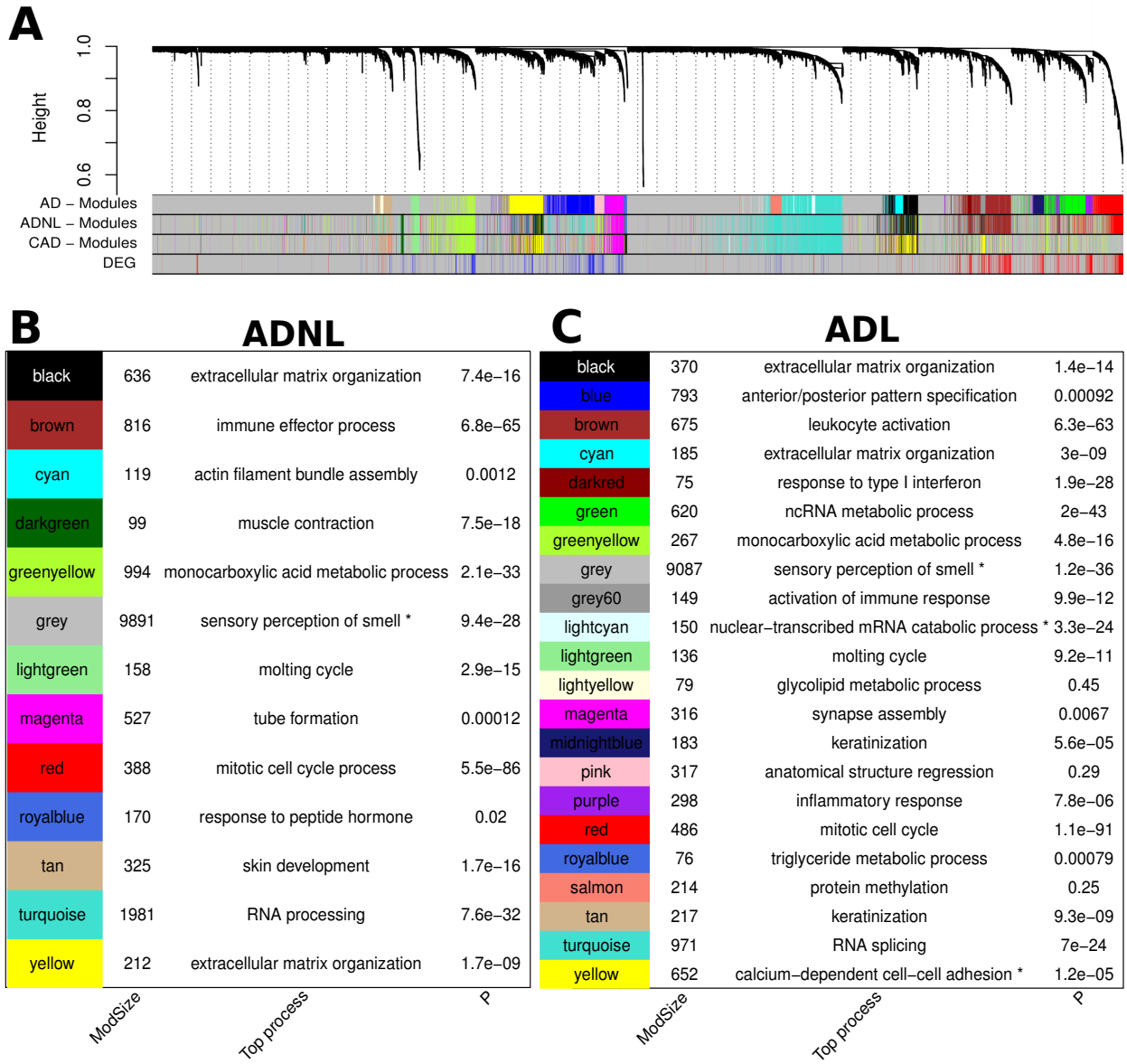

Figure 6.2: Module definitions of AD co-expression networks, gene tree and enrichment analysis. (A) Hierarchical clustering of TOM dissimilarity and module assignments. Tracks below the dendrogram correspond to module definitions in the ADL, ADNL and CAD networks as well as differential status ($p < 0.05$, $LFC > 0.58$). (B) Module size and top GO biological process term for each module in the ADNL network. (C) Module size and GO terms for the ADL network.

for key skin associated functions.

First, considering the control network (CAD), 10 modules were identified and all subnetworks were over-represented for a GO BP ($p < 0.05$, **Appendix B, Figure B.3**). Two modules were associated with skin development processes (salmon, lightgreen), and one was associated with the extracellular matrix (yellow). Housekeeping terms such as 'RNA processing' and 'monocarbocylic acid metabolic process' were highly enriched. There were no modules which were enriched for immune system processes suggesting that co-expression relationships between immune genes in a state of homeostasis are less defined.

6.3.3.2 ADNL network modules

Regarding ADNL, 13 modules were identified, and all were enriched for a GO biological process ($p < 0.05$, **Figure 6.2 B**). One module was enriched for immune processes (brown), three modules were enriched for the extracellular matrix (black, yellow, cyan) and two modules were associated with skin specific development processes (tan, lightgreen). Several modules were enriched for housekeeping functions such as 'RNA processing' (turquoise) and 'mitotic cell cycle process' (red).

6.3.3.3 ADL network modules

Next, GO analysis of lesional AD modules was performed (**Figure 6.2 C**). Of the 22 ADL modules identified, 19 were enriched for a GO biological process ($p < 0.05$). Three modules were strongly enriched for immune system processes. The brown module was enriched for 'Leukocyte activation' ($p = 6.3e\text{-}63$), the darkred module was enriched for 'response to type I interferon' ($p = 1.9e\text{-}28$) and the purple module was enriched for 'inflammatory response' ($p = 7.8e\text{-}06$) demonstrating that co-expression amongst genes encoding immunity is a key component in AD. As observed in ADNL, three modules were enriched for the ECM (black, cyan, yellow) and three modules were associated with skin development processes such as lightgreen which was enriched for 'molting cycle' ($p = 9.2e\text{-}11$), darkblue which was enriched for 'keratinization' ($p = 5.6e\text{-}05$) and the tan module which was also enriched for 'keratinization' ($p = 9.3e\text{-}09$). A complete list of the top GO terms enriched for each module is shown in (**Appendix B, Tables B.1 - B.6**).

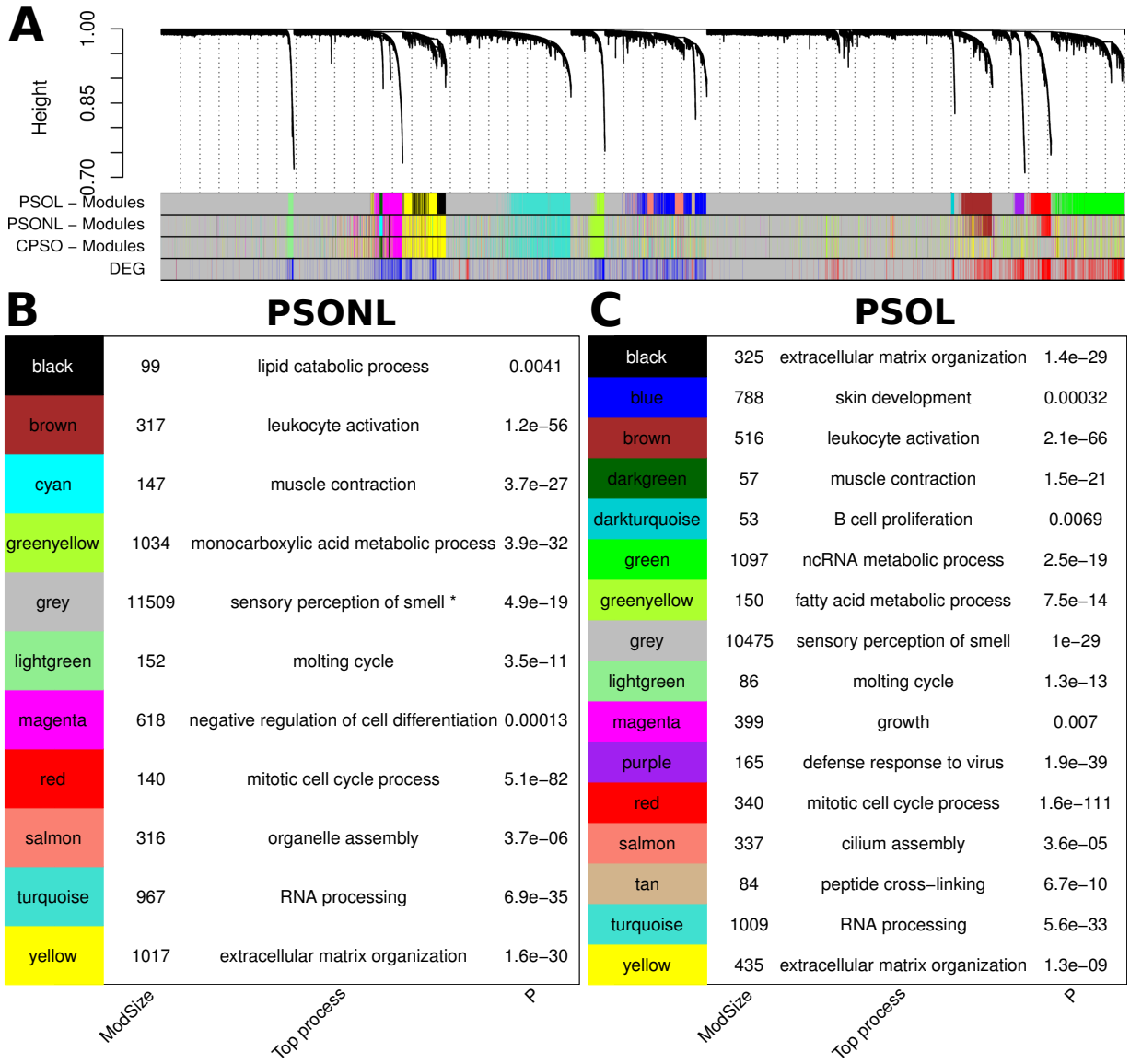

Figure 6.3: Module definitions of PSO co-expression networks, gene tree and enrichment analysis. (A) Hierarchical clustering of TOM dissimilarity and module assignments. Tracks below the dendrogram correspond to module definitions in the PSOL, PSONL and CPSO networks as well as differential status ($p < 0.05$, $LFC > 0.58$). (B) Module size and top GO biological process term for each module in the PSONL network. (C) Module size and GO terms for the PSOL network.

6.3.4 Psoriasis associated networks

The psoriasis co-expression networks were next analysed in terms of module composition. The PSOL, PSONL and CPSO networks contained 16, 11 and 10 modules respectively (**Table 6.2, Figure 6.3 A**). As observed in AD, a qualitative analysis of module definitions as presented in (**Figure 6.3 A**) showed a general concordance across networks indicating that there may be a degree of preservation in co-expression relationships.

6.3.4.1 CPSO network modules

To determine the biological function of co-expression modules, each module was subjected to gene ontology enrichment analysis. Ten modules were identified in the PSO control network (CPSO), all of which were over-represented with a GO BP term ($p < 0.05$, **Appendix B, Figure B.3**). Two modules were associated with skin development processes (tan, lightgreen), and one was associated with the extracellular matrix (yellow). Several modules were enriched for housekeeping terms such as 'RNA processing' (turquoise), 'inorganic ion transmembrane transport' (magenta) and 'monocarboxylic acid metabolic process' (greenyellow). There were no modules which were enriched for immunity, demonstrating state of homeostasis in healthy samples.

6.3.4.2 PSONL network modules

Next, modules within the PSONL network were subjected to GO analysis (**Figure 6.3 B**). All modules were enriched for a GO BP term ($p < 0.05$), and like in AD, many of the functions resembled core biological processes in the skin. The brown module was highly enriched for 'leukocyte activation' ($p = 1.2e\text{-}56$) indicating that co-expression of immune processes is well defined, even in non-lesional skin. Other key modules included the lightgreen module which was was enriched for 'molting cycle' ($p = 3.5e\text{-}11$) and the yellow module was enriched for 'extracellular matrix organisation' ($p = 1.6e\text{-}30$).

6.3.4.3 PSOL network modules

All lesional sub networks in PSO were enriched for a GO BP ($p < 0.05$, **Figure 6.3 C**). Enrichment results revealed that three modules were enriched for immune system processes including the brown module which was enriched 'leukocyte activation' ($p = 2.1e\text{-}66$), the darkturquoise module which was enriched for 'B cell proliferation' ($p = 0.0069$) indicating a potential role for B cells in PSO, and the purple module which was highly enriched for

'defence response to virus' (p = 1.9e-39). Two modules were enriched for ECM processes (black, cyan), and two modules were enriched for skin associated processes (lightgreen, blue).

Overall, a lack of enrichment for immune system proccess in control networks was observed reflecting a state of homeostatic equilibrium. In comparison, immune system modules were well defined in disease associated networks, even in a non-lesional state. This general overview of the reconstructed networks revealed co-expression modules which were highly relevant to skin disease and provide a basis for further analysis into the similarity and differences in transcriptional architecture across cohorts, as well as the association of subnetworks with microbial abundance.

6.3.5 Inflammatory network module preservation

In **Chapter 4**, differences in transcript expression were quantified by means of differential gene expression analysis, however, this approach did not consider the relationship between genes. A module preservation analysis was performed to determine the consistency and robustness of gene-gene interactions across networks. Module preservation analysis allows evaluation of the dynamic properties of gene modules and enables study into how community structure changes in the context of disease [232]. In contrast to differential analysis, the main objective of module preservation analysis is to to identify co-expression modules which are disrupted indicating a potential rewiring of gene-gene interactions occurring between healthy, non-lesional and lesional status.

The preservation of modules across networks was analysed with two main strategies (see **Section 6.2.5** for a detailed discussion). The first, and most naive approach is the application of cross tabulation statistics, which determines the significance of overlap in the membership of gene modules between networks and does not take into account information regarding the patterns of gene-gene interactions. The second strategy corresponds to network based preservation statistics. As defined in [232], network based statistics relate to a plethora of strategies designed to determine whether the connectivity and density patterns within a reference network are preserved in a designated test network.

6.3.5.1 Cross tabulation statistics for module preservation

Cross tabulation statistics were applied to obtain a general overview into the concordance of module definitions between healthy, uninvolved and lesional networks. Regarding AD, a remarkable degree of preservation was observed between the ADL, ADNL and CAD networks where each network module had at least one significant overlap with another module ($p < 1e\text{-}3$, **Appendix B, Figure B.4**). A similar trend was observed in PSO where the vast majority of modules were highly conserved (**Appendix B, Figure B.5**). The cross-tabulation statistics were also applied between the ADL and PSOL networks which again demonstrated a similar transcriptional architecture between these diseases (**Appendix B, Figure B.6**).

Overall, a strong level of module preservation was observed across networks, however qualitatively, there were indications that a small number of modules may only by weakly preserved. For example, the blue and green modules in the PSOL network appeared to be split across several PSONL and CAD modules (**Appendix B, Figure B.5**). Cross tabulation statistics are sensitive to the parameters which define the network as well as the community detection strategy [232] and do not take into account correlation strengths between module members, therefore, further analysis of module preservation was performed using network based statistics.

6.3.5.2 Preservation of network modules in AD

First, ADL was used as the reference network and preservation statistics were calculated to determine if ADL modules could be identified within the ADNL adjacency matrix. Using the guidelines suggested by [232], a $Z_{summary}$ threshold of < 10 was used to denote weak preservation and $Z > 10$ reflected strong module preservation. As observed with the cross-tabulation analysis, a remarkable degree of preservation of ADL modules was observed (**Figure 6.4 A**). All of the ADL modules were preserved with a $Z_{summary} > 10$. When considering the CAD network as the test network, and ADL as the reference network, the green and purple modules were only weakly preserved ($Z < 10$, **Figure 6.4 B**). Whilst the $Z_{summary}$ statistic suggested a high degree of preservation, Zconnectivity in our networks was strongly correlated with module size (**Appendix B, Figure B.7**). This is a known limitation of the method, therefore, the $Z_{summary}$ results were supplemented with the medianRank statistic which is a robust way of reporting module preservation without the influence of module size.

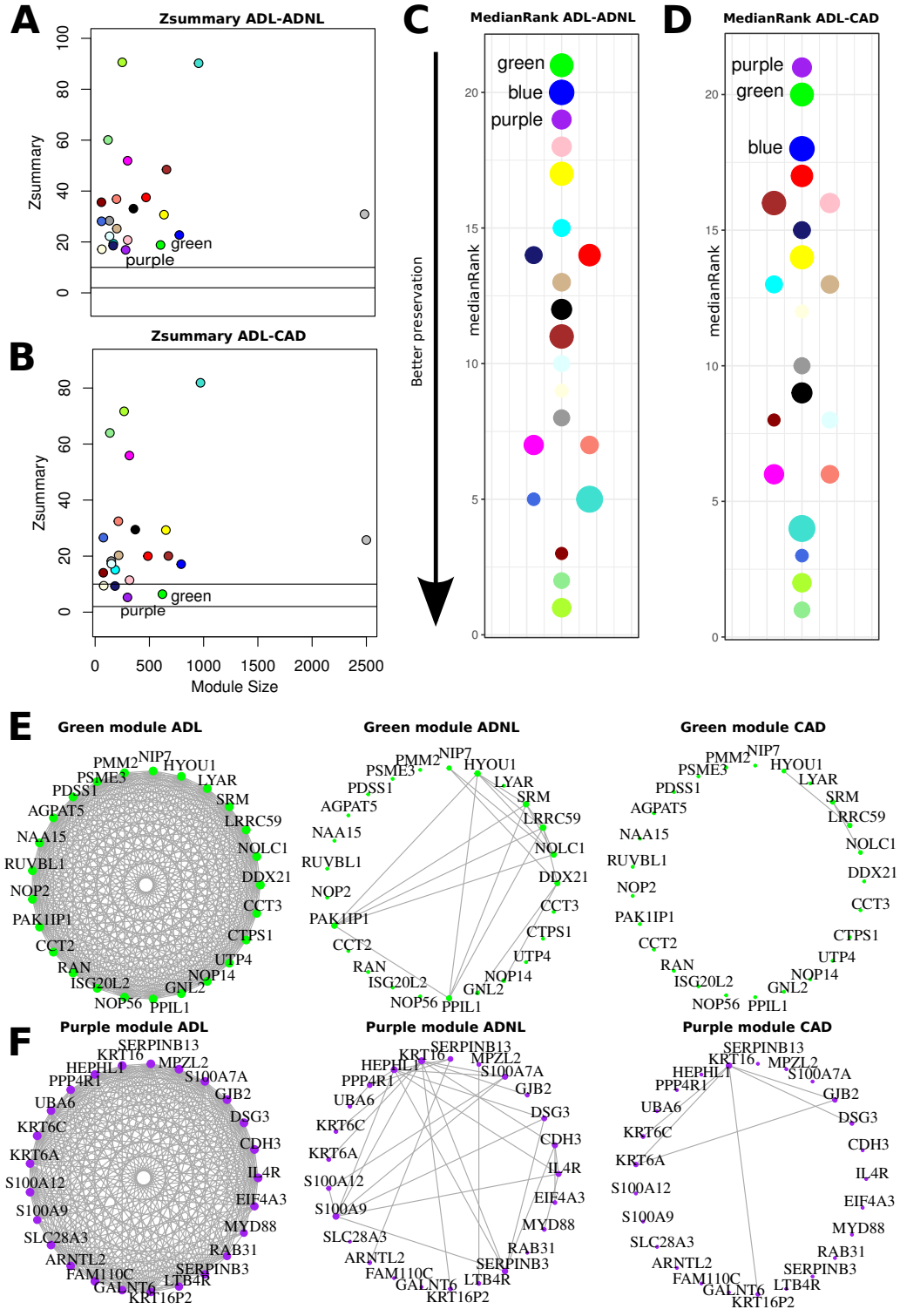

Figure 6.4: Module preservation analysis of ADL modules. (A) $Z_{summary}$ of modules using ADL modules as reference and the ADNL expression matrix as the test set. (B) Z summary of modules with ADL reference and CAD test set. (C-D) Median rank statistics, modules ranked closer to 1 are highly preserved, and modules with high ranks are weakly preserved. (C) MedianRank statistics with ADL reference and ADNL test. (D) ADL reference and ADNL test. (E-F) Network diagrams depicting the correlations between hub genes of weakly preserved modules. An edge represents a correlation > 0.65 in the ADL, ADNL and CAD networks. (E) green module, (F) purple module.

Analysis of module preservation statistics using the medianRank scoring method indicated that the green module was the most weakly preserved within the ADNL network (**Figure 6.4 C)**, and that the purple and green modules were the most weakly preserved within the CAD network (**Figure 6.4 D)**. Further investigation of these modules showed that the connectivity patterns amongst the hub genes in the green module were severely depleted in the ADNL and CAD networks (**Figure 6.4 E)**. The top GO term for the green module was for ncRNA metabolic process (p =1.97e-43, **Figure 6.2 C**) and the enriched pathways reflect processes involved in protein and RNA processing (**Appendix B, Figure B.1**). A similar pattern was observed within the purple module (**Figure 6.4 F)** which was enriched for 'inflammatory response' (p = 7.8e-06) of which the Th2 cytokine receptor IL4R was a hub gene, and contained key immune genes as well as antimicrobial peptides. Given that the connectivity patterns amongst the green and purple modules were significantly perturbed, these modules likely represent AD associated subnetworks.

6.3.5.3 Preservation of network modules in PSO

The same approach was applied using PSOL as the reference network, and the PSONL and CPSO networks as the test sets. The vast majority of modules were preserved except the green module which had a $Z_{summary}$ statistic lower than 10 in both the PSONL and CAD networks (**Figure 6.5 A)**. No other modules had a $Z_{summary}$ statistic lower than 10. Given that the $Z_{summary}$ statistic is sensitive to differences in module size, the results were supplemented with the medianRank. This analysis also indicated that the green module was the least preserved module across networks (**Figure 6.5 B-C**). A complete breakdown in the correlations amongst the top hub genes was observed across networks (**Figure 6.5 D**) demonstrating that interactions within this module undergo significant rewiring during inflammation. The green module was significantly enriched for ncRNA metabolic process (p = 2.5e-19, **Figure 6.3 C**) which has recently been associated with psoriatic inflammation [243]. Interestingly, the cytokine receptor IL4R was also amongst the hub genes of this module suggesting that this subnetwork may also represent inflammatory component. IL4R was found to be a hub gene within the ADL-purple module which was only weakly preserved across AD networks. The PSOL green module was strongly enriched for genes originating from both the green and purple ADL modules (**Appendix B, Figure B.6**) which could indicate a relationship between inflammation and ncRNA processes.

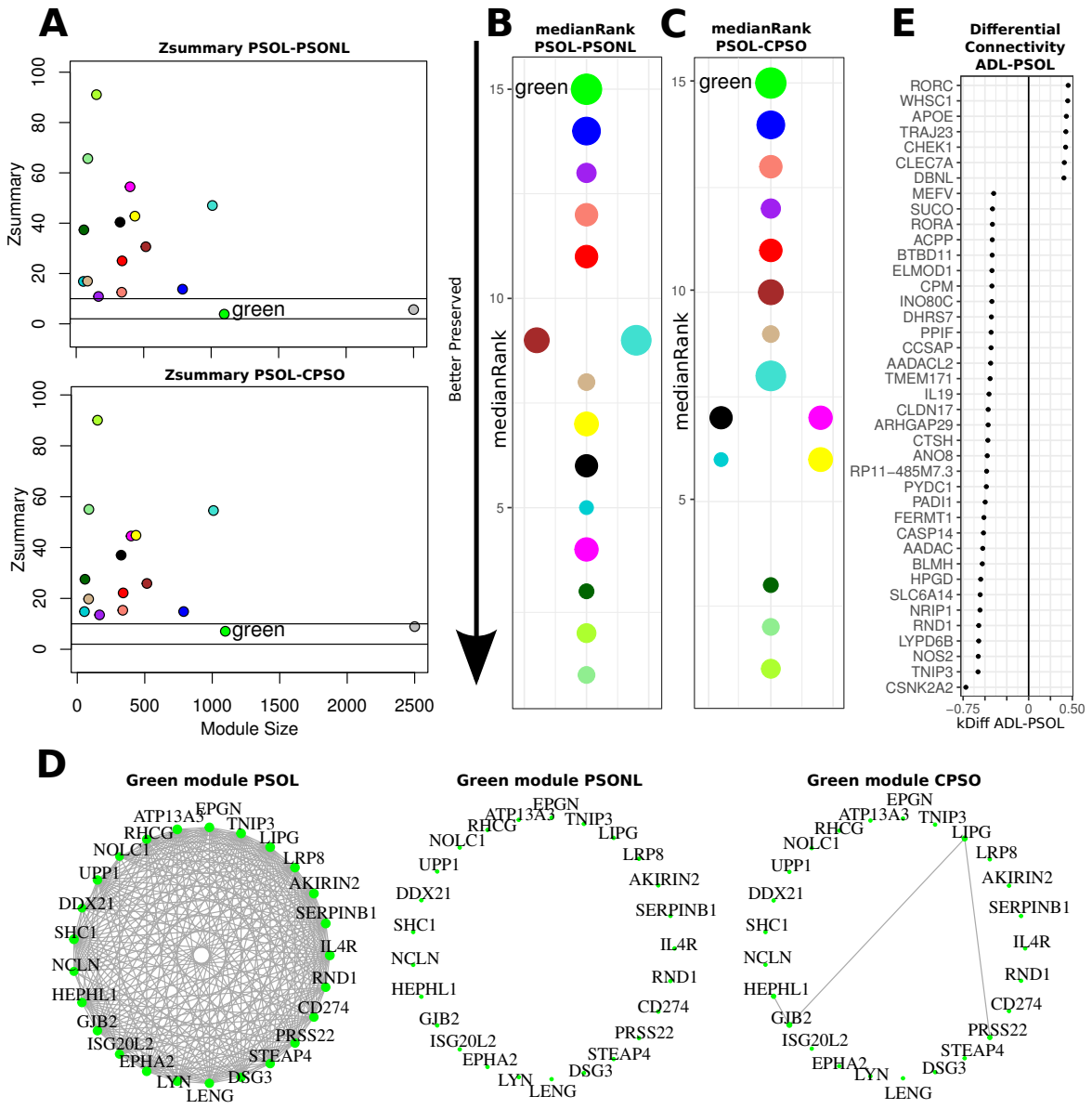

Figure 6.5: Module preservation analysis of PSOL modules. (A) $Z_{summary}$ of modules using PSOL modules as reference and the PSONL expression matrix as the test set (top), and Z summary of modules with PSOL reference and CPSO test set (bottom). (B-C) Median rank statistics. Modules ranked closer to 1 are highly preserved, and modules with high ranks are weakly preserved. (B) MedianRank statistics with PSOL reference and PSONL test. (C) PSOL reference and CPSO test. (D) Network diagrams depicting the correlations between hub genes of the green module. An edge represents a correlation > 0.65 in the PSOL, PSONL and CPSO networks. (E) Differential connectivity analysis of genes for the ADL-PSOL comparison. kDiff was calculated as whole network connectivities in the PSOL network subtracted from connectivity in the ADL network kADL-kPSOL. Genes with kDiff > 0.4 are shown.

In attempt to determine if differentially co-expressed modules were present between the PSOL and ADL networks, the module preservation analysis was performed using the ADL network as the reference network and PSOL as the test network. The same procedure was performed using PSOL as the reference and ADL as the test network. In both cases, all modules were highly preserved $Z_{summary} > 20$ (data not shown) demonstrating a strong relationship between the co-expression relationships which underlie inflammation in the skin. Whilst module structure was preserved between ADL and PSOL, individual genes could display different connectivity patterns which may provide insight into the master transcriptional regulators which underlie the Atopic and Psoriatic systems. A differential connectivity analysis was performed between ADL and PSOL by calculating the difference in network connectivity across all genes. That is, in the context of lesional disease, the overall weighted connectivity of a gene with other nodes in the network is calculated independently for both networks; kA and kP. For each gene the connectivity difference, $kA - kP$, is calculated [151]. Seven genes were of greater connectivity in AD and 33 genes in PSO (kDiff > 0.4, **Figure 6.5 E**). Genes differentially connected in AD related to genes associated with T cells such as the differentiation regulator RORC, as well as TRAJ23 and CLEC7A. The cholesterol transporter Apolipoprotein E (APOE) was also differentially connected in AD. Three genes of greater connectivity in PSOL were associated with negative regulation of NFKB signalling including NOS2, TNIP3 and PYDC1.

6.3.6 Association of co-expression modules with the microbiome

The expression profile of a co-expression network module can be summarised into a single representative vector known as the module eigengene. This module eigengene can then be used to link the expression of a module to clinical traits. In this section the relationship between eigengene expression and microbial abundance as well as disease severity (SCORAD in AD, PASI in PSO) was evaluated in attempt to unearth potential host-microbe associations of clinical relevance.

Only associations between module eigengenes and traits were significant within the ADL network (**Appendix B, Figure B.8)**. Regarding the lesional ADL network, 6 modules were significantly associated with microbial abundance, two of which were also significantly associated with SCORAD (**Figure 6.6 A**). Regarding the microbiota, only associations between *Staphylococcus aureus* and co-expression modules were identified, in line with previous results presented in **Chapter 5**.

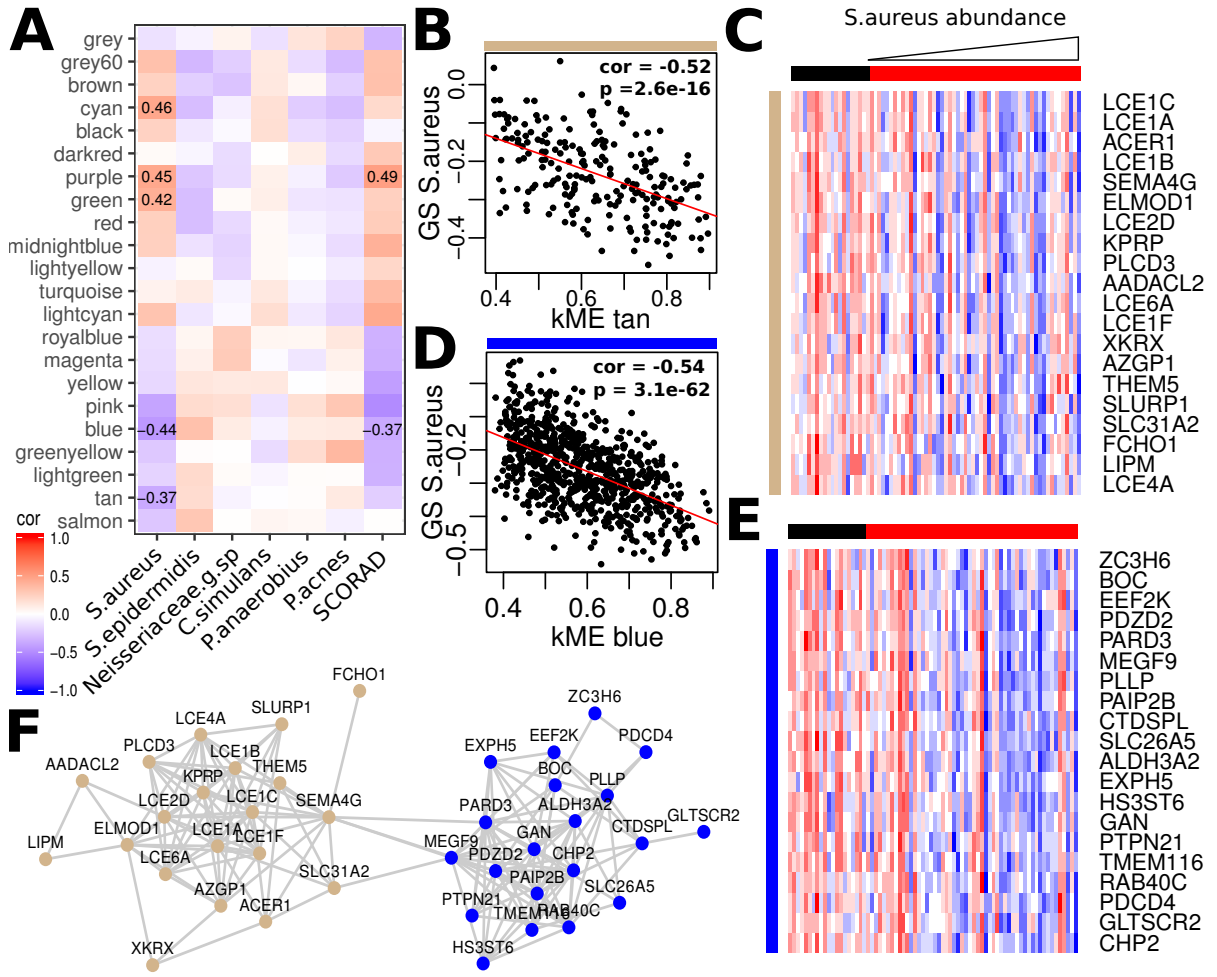

Figure 6.6: Associations between co-expression modules, disease severity and microbial abundance in AD. (A) Association between module eigengenes, microbes and SCORAD. Significant associations with an adjusted p value < 0.1 and $r > 0.3$ are shown. (B) Pearson correlation between kME and *S. aureus* gene significance in the tan module. (C) Heatmap of the top 20 hub genes ranked by kME in the tan modules. Samples are ordered by increasing *S. aureus* abundance. Black column labels correspond to *S. aureus* negative, and red labels correspond to *S. aureus* positive. (D) kME vs gene significance for blue module. (E) heatmap of hubs in the blue module. (F) Network visualisation of the tan and blue modules. The top 10% of correlations between the top 20 hub genes in the tan and blue modules.

6.3.6.1 Tan module

First, negative relationships were investigated of which the tan module was inversely correlated with *S. aureus* (r = -0.37, p = 0.015, **Figure 6.6 A**). This module was was strongly enriched for GO processes associated with skin development such as keratinization, keratinocyte differentiation, and epidermal cell differentiation (p < 9.5e-09, **Appendix B, Table B.1**). To identify potential hub genes within this module, module membership values (kME) were calculated for each gene as the correlation between the gene member and module eigengene. A measure of gene significance for *S. aureus* was also calculated as the correlation between the gene and *S. aureus*. The correlation between kME and gene significance showed a highly significant inverse relationship indicating that key genes within the tan module are negatively correlated with *S. aureus* abundance (**Figure 6.6 B**). Genes were then ranked by kME value to identify hub genes, and the top ranking genes were selected and presented in a heatmap whereby samples were ordered by *S. aureus* abundance (**Figure 6.6 C**). The heatmap represents a pattern by which hub genes of the tan module are of lower expression in the presence of high *S. aureus* relative abundance.

The top genes within the tan module clearly reflect genes associated with the epidermal differentiation complex. Several members of the late cornified envelope family (LCE) including LCE1C, LCE1A, LCE1B, LCE2D, LCE6A, LCE1F and LCE4A, as well as ACER1, and KPRP were amongst the top hub genes. The tan module also included other key components of the epidermal barrier which were highly ranked including FLG2, FLG and Loricrin (LOR). Barrier defect is known to be a characteristic of AD [44], and is specifically associated with mutations with the FLG gene [50], therefore, this finding demonstrates that transcriptional activity of the skin barrier is associated with the colonisation of pathogens such as *S. aureus*.

6.3.6.2 Blue module

The blue module was negatively associated with *S. aureus* abundance (r = -0.44, p < 0.05, **Figure 6.6 A)**. The top GO terms associated with this module were anterior/posterior pattern specification and regionalization, however, the significance of enrichment for the blue module was one of the lowest across the whole ADL network indicating that the genes in this module are likely to be heterogeneous and may represent currently unknown processes (p < 9e-04 **Appendix B, Table B.1**). *S. aureus* gene significance and kME for blue module members were correlated and highly significant (r = -0.54, p = 3.1e-62,

Figure 6.6 D) indicating that hub genes are associated with *S. aureus* abundance.

Top hub genes included several candidate genes identified in **Chapter 5** including GAN, PARD3, HS3ST6, TMEM116, BOC, CHP2 (**Figure 6.6 E**). Other key genes identified in **Chapter 5** were also present in the blue module and highly ranked such as IL34 (rank = 45) and RORC (rank = 34) strengthening the relationship between these genes and association with *S. aureus* in ADL. An inverse association with disease severity ($p < 0.05$, **Figure 6.6 A**) was observed, therefore, the genes in this module could be of clinical relevance and further study could elucidate their function in the pathogenesis of AD.

6.3.6.3 Green module

One of the larger modules with 620 genes in the ADL network, green, was strongly enriched for the GO term 'ncRNA metabolic process' (p = 1.97e-43, **Appendix B, Table B.1**) and pathways such as 'Ribosome biogenesis in eukaryotes'. A significant association between *S. aureus* and the green module was found (**Figure 6.6 A)** and the top hub genes were of greater expression in the presence of *S. aureus* (**Figure 6.7 A-B)**. There is an increasing appreciation for the relationship between non-coding RNAs and inflammation [244], and psoriasis has recently been associated with long non-coding RNAs [243], although the mechanisms are poorly understood. Several of the top hub genes were associated with ribosomal RNA (**Figure 6.7 B-C**) which is traditionally considered to play a housekeeping role. Whilst the interpretation of this module in the context of AD is challenging, further investigations into the role of ncRNA, protein metabolism and the association with *S. aureus* could provide insight into this relationship.

6.3.6.4 Purple module

The purple module was an immune system related subnetwork and was enriched for 'inflammatory response' and 'innate immune response' (p < 5e-05, **Appendix B, Table B.1**). This module significantly correlated with *S. aureus* ($r = 0.45$, p = 0.018, **Figure 6.6 A**) and module membership showed a positive trend with GS ($r = 0.56$, p = 2.2e-26, **Figure 6.7 D**). Many of the top hub genes within the purple module were associated with antimicrobial defence including S100A9, S100A12 and S100A7A and also included the Th2 cytokine receptor IL4R (**Figure 6.7 E-F**). DEFB4A was also highly ranked by kME (rank = 39). As well genes encoding for immunity, other hubs included KRT6C, KRT16, DSG3, GJB2 and the serine protease SERPINB13 which are critical for structural

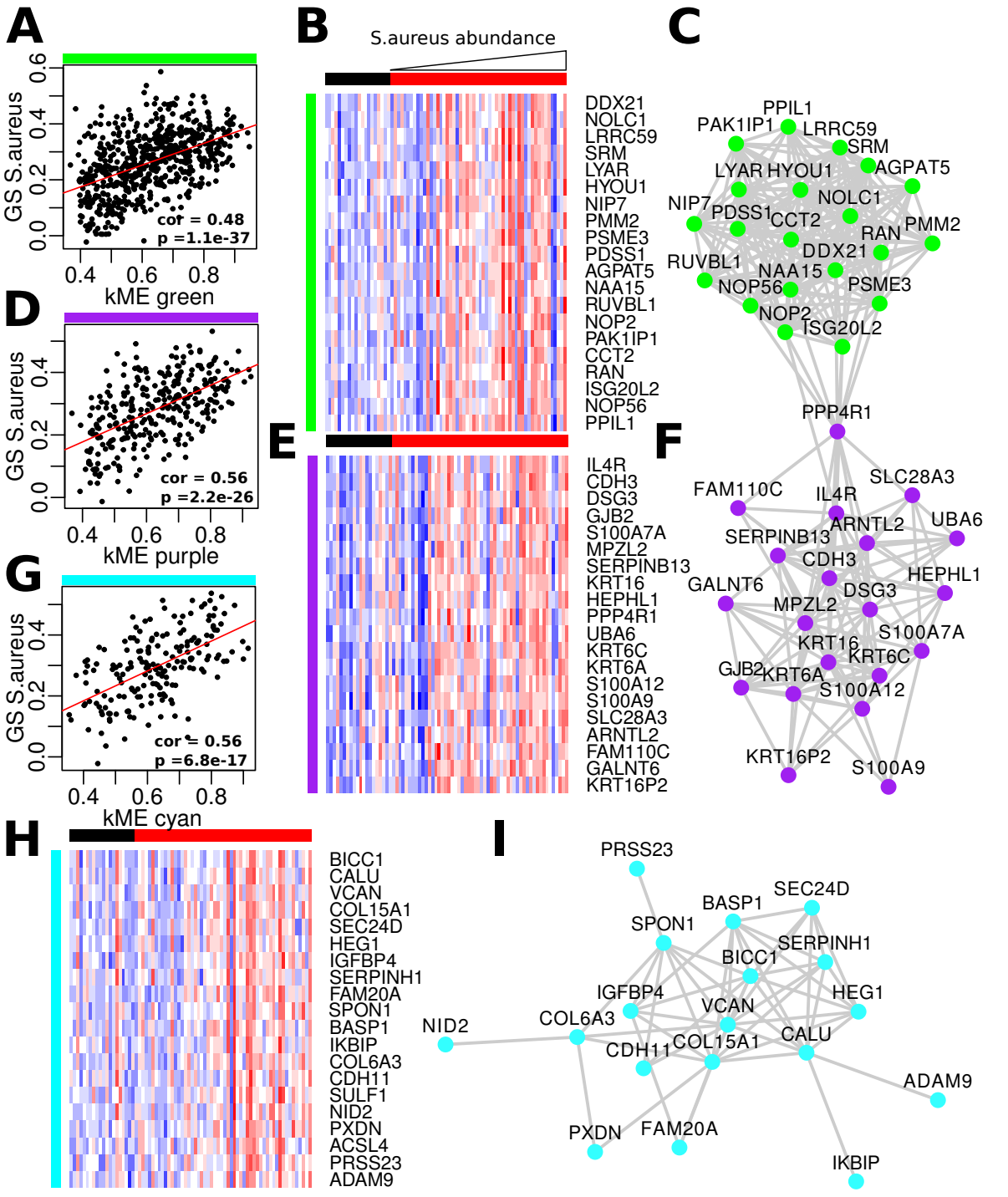

Figure 6.7: Co-expression modules positively associated with *S. aureus*.(A) kME vs S. aureus gene significance for the green module. (B) Heatmap of the top 20 hub genes ranked by kME in the green module. Samples are ordered by increasing *S. aureus* abundance. Black column labels correspond to *S. aureus* negative, and red labels correspond to *S. aureus* positive. (C) Network diagram of the top hubs in the green module. The top 10 % of correlations between the top 20 hub genes in green module are shown. (D) kME vs gene significance for the purple module. (E) heatmap of hubs in the purple module. (F) network diagram in the purple module. (G) kME vs gene significance for the cyan module. (H) heatmap of hubs in the cyan module. (I) Network diagram for the cyan module.

integrity and homeostasis of the epidermal barrier.

Pathway analysis of module members revealed further enrichment for innate pathways including 'NF-κB signalling' and 'cytosolic sensors of pathogen-associated DNA' ($p < 0.01$, **Figure 6.8**), therefore, this module represents components of the innate immune system and epidermal barrier which are active in the presence of *S. aureus*. As this module was significantly associated with disease severity ($p = 0.015$, **Figure 6.6 A**), it clinically is a relevant module and could represent a crosstalk between the immune system and epidermal barrier and may provide insights into host response to *S. aureus* infection in AD.

6.3.6.5 Cyan module

The cyan module was strongly enriched for GO terms associated with the extracellular matrix such as 'extracellular matrix organisation' and 'extracellular structure organization' ($p <$ 3e-9, **Figure 6.2 C**, **Appendix B, Table B.1**). The top enriched pathways also reflected ECM association with enrichment of 'ECM-receptor interaction' and 'collagen biosynthesis' (**Figure 6.8**). A significant association between the cyan module eigengene and *S. aureus* ($r = 0.46$, $p = 0.015$, **Figure 6.6 A**) as well as a positive (non-significant) trend for disease severity was observed. As with other *S. aureus* associated modules, kME and GS were significantly correlated ($r = 0.56$, p = 6.8e-17, **Figure 6.7 G**).

Genes within the cyan module (**Figure 6.7 H-I**) included the collagen family proteins COL15A1 and COL6A3. Several other collagen family genes were highly ranked within the cyan module including COL4A1, COL4A2, COL6A2 and COL6A1. Other ECM genes were amongst the top hubs including VCAN which is involved in cell-cell adhesion, NID2, which binds collagen and is associated with the structure of the basement membrane, and the serine proteinase inhibitor SERPINH1. Several of the ADAM family of disintegrin and metalloproteases were present in the cyan module, including ADAM9, ADAMTS12 and ADAM12 which are involved in wound healing and possess anti-angiogenic properties [245, 246]. The ECM plays an important role within inflamed tissues and influences immune cell signalling, migration, activation and T cell polarization [214]. The association of this module with *S. aureus* establishes a link between between the ECM and pathogen abundance which could relate to the remodelling and fibrosis which occurs in atopic lesions [208].

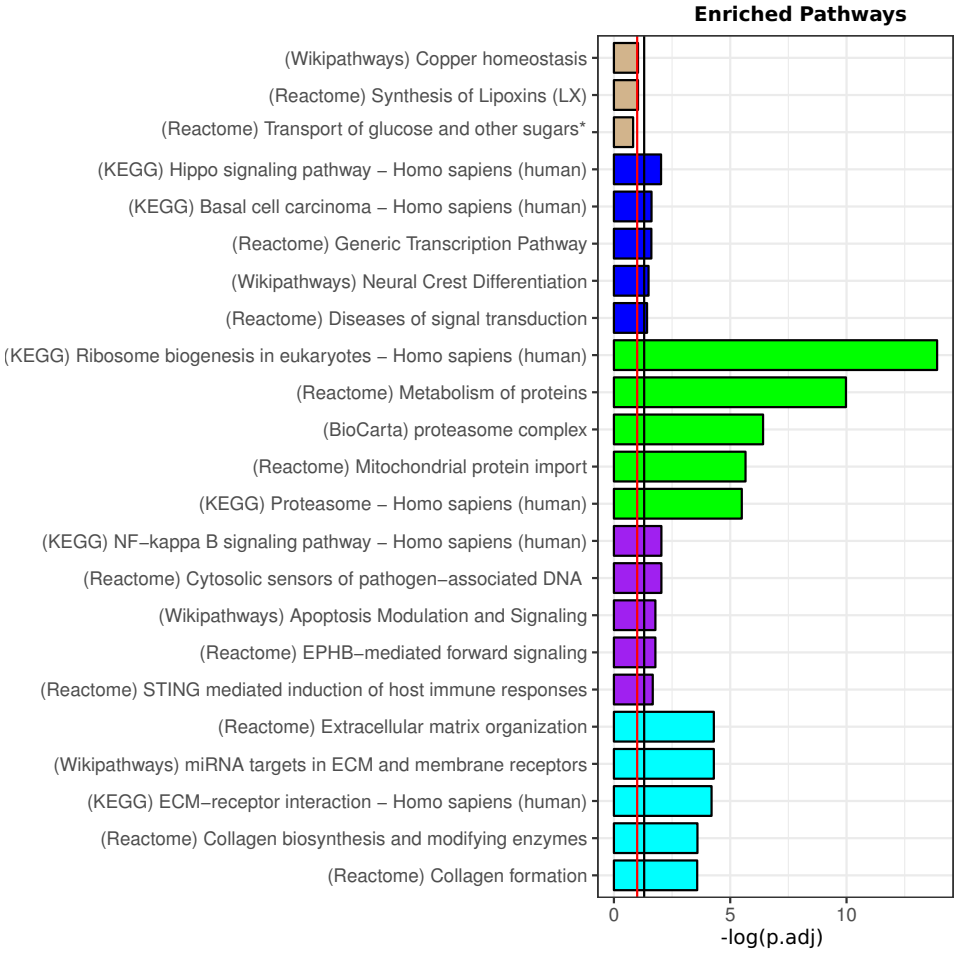

Figure 6.8: Top 5 enriched pathways for modules of interest as identified by consensusPathDB. The black line corresponds to p = 0.05, and the red line to p = 0.1.

6.4 Conclusions and Discussion

In the analysis presented, independent co-expression networks were constructed to investigate the complex systems underlying skin inflammation. These co-expression networks allowed for investigation into core sets of genes which interact during atopic and psoriatic inflammation. Furthermore, dimensionality reduction via the calculation of module eigengenes facilitated the mapping of host-microbe interactions identifying gene sets related to pathogen abundance, and enabled the identification of those which covaried with disease severity. The vast majority of modules identified across networks were strongly enriched for biological processes relevant to skin biology.

First, a module preservation analysis was performed, and despite the considerable expression differences described in **Chapter 4**, it was surprising to find that the transcriptional architecture was remarkably preserved across networks. Despite an overall state of preservation, two modules were found to be weakly preserved in AD and one module in PSO compared to controls and their non-lesional counterparts. One of these modules, the green module, was highly enriched for non coding RNA metabolic processes as well as 'ribosome biogenesis' (in both PSO and AD). Recent studies in psoriasis have shown that lncRNAs were significantly enriched amongst psoriasis associated modules and could play important roles in the regulation of inflammatory pathways [243]. In ADL, the green module was also associated with *S. aureus* abundance and indicates an interesting relationship between *S. aureus* abundance and housekeeping RNA processes.

In atopic dermatitis, the purple module was associated with inflammatory response which is not preserved in either of the ADNL or CAD networks. The top hub gene of the purple module was IL4R, and whilst it is well established that genes associated with Th2 cells are involved in atopic inflammation [44], this analysis demonstrated that genes within this module form strong co-expression relationships during inflammation and are thus differentially co-expressed. As IL4R was the top hub gene of this module, is it likely that the purple module is associated with increased Th2 activity. Furthermore, this module was associated with disease severity and *Staphylococcus aureus* abundance indicating that Th2 activity exacerbates inflammation and is related to pathogen colonisation in AD. It was interesting to find that all modules were highly preserved between ADL and PSOL indicating that the transcriptional architecture of these diseases is similar. Despite this, several genes were found to be differentially connected between diseases and may act as differential regulators.

One of the main findings of this analysis is that only associations between *Staphylococcus aureus* in the ADL network were identified. In **Chapter 3** uninvolved atopic skin was shown to be dominated by *S. aureus*, yet no significant associations between *S. aureus* in uninvolved networks were identified. This suggests that *S. aureus* covaries with inflammation yet it remains unclear how or if *S. aureus* triggers a flare. This observation could reflect findings of higher abundances of staphylococcal enterotoxin isolated from *S. aureus* strains on lesional skin than non-lesional skin [66]. Despite expansion in specific microbes in PSO, no significant associations were identified between potential pathogens and transcriptional activity in PSO.

Analysis of the ADL network showed a negative association between the tan module and *S. aureus* abundance. This module contained a number of genes responsible for the structural integrity of the epidermal barrier and consisted of FLG, FLG2, LOR and LCE family genes. Within the skin, the epidermis harbours a diverse range of specialised immune cells including langerhans and dendritic epidermal T cells which along with keratinocytes, act as the first line of defence against the external environment and invading pathogens [160]. It is well established that AD is associated with epidermal barrier deficiencies [51, 204], the most well characterised being loss of function mutations within the FLG gene [49, 50] which aggregates keratin filaments in the epidermis and is a key component of the stratum corneum. Breakdown products of FLG are used to generate natural moisturising factors which play roles in hydration of the skin barrier. FLG loss of function mutations are therefore related to the dryness often characteristic of AD lesions [49]. Studies have shown that barrier defects are associated with increased cutaneous infiltration of antigens, thus, the increased permeability of the epidermal barrier could be a key component of atopic inflammation [52]. The association of the tan module and *S. aureus* suggests that disruption to barrier integrity through reduced expression of this module is related to increased abundance of pathogen.

As well as the epidermal compartment, the strongest host-*S. aureus* association involved a module encoding for components of the ECM which makes up a large proportion of the dermis. The cyan module represented a component of the ECM which was specifically perturbed in the presence of increased *S. aureus* abundance. Tissue and extracellular matrix remodelling is characteristic of atopic inflammation [208] and *Staphylococcus aureus* has

been shown to induce MMP expression in vitro in fibroblasts [247], therefore, it is possible that *S. aureus* could be associated with ECM remodelling. Cytokines produced by Treg cells such as TGFB and IL10 are upregulated in AD (see **Chapter 3, Figure 4.7**). TGFB is an immunosuppresive cytokine which can induce fibrosis [248] and can be induced by the Th2 cytokine IL13 [198]. Furthermore, Tregs are thought to possess both pro and antifibrotic properties [248], thus *S. aureus* may be associated with a Th2 response [249] which indirectly induces ECM remodelling.

S. aureus was also negatively correlated with the blue module and although enrichment analysis was weak and inconclusive, it suggests that this module contains genes of unknown function in the context of AD. Interesting candidates within this module included IL34 which was also found to be associated with *S. aureus* in **Chapter 5** and has recently been associated with AD where it is thought to inhibit inflammatory cascades [211]. As this module was also inversely correlated with disease severity, further analysis could be performed to deduce the function of this potentially clinically relevant module.

In summary, a large scale integrative network analysis of the transcriptomic processes underlying psoriatic and atopic inflammation was performed. Several co-expression modules were related to *S. aureus* abundance and clinical severity in AD. Furthermore, although not significant, in modules which were positively correlated with *S. aureus*, a trend for increased disease severity was observed indicating that *S. aureus* exacerbates inflammatory response. The co-expression modules identified were highly relevant to structural and inflammatory processes underlying inflammation and further experimental analysis can continue to unravel the complex interactions between pathogens and to identify potential therapeutic targets.

Chapter 7

Conclusions and future perspectives

Advances in sequencing technologies have revolutionised our understanding of the resident flora which envelop the surfaces of our bodies. In recent years, many 16S and metagenomics surveys have been performed and it is now becoming increasingly clear that the relationship between microbiota and host health is not as simple as once perceived. Many analyses of the microbiota colonising the surface of diseased tissues have reported an apparent state of dysbiosis which represents broad shifts in community composition. Collectively, these analyses have shown that the relationship between the microbiome and host health is complex, however, much of the evidence suggests that the resident microbiota and the host immune system exists in delicate homeostatic balance which when disturbed, can provoke an inappropriate immune response [31, 22]. Further work is still required to determine the principles, however, as the cost of sequencing drops and the bioinformatics methods to analyse the vast quantities of challenging data improve, greater understanding of host microbiota interactions may enable modulation of the resident microbiota for therapeutic applications.

Collectively the results presented in this thesis support the notion of a complex ecosystem of microbiota colonising the healthy and inflamed cutaneous membrane. The largest survey to date of the inflamed microbiota is presented in **Chapter 3** which revealed the major compositional differences between atopic dermatitis, psoriasis and healthy skin. This analysis portrayed a diverse cutaneous microbiota with over 4000 unique OTUs present on healthy skin, with each site consisting of approximately 143 different taxa.

Clear compositional differences were observed between disease types, and dysbiosis is undoubtedly a major factor of the AD microbiome. Compositional shifts in psoriasis were more subtle, however, in both conditions severe disease was characteristic of reduced diversity thus, demonstrating a relationship between homestatic balance and cutaneous inflammation. It has been suggested that *S. aureus* may be driving dysbiosis in AD [64], and the co-occurrence network analysis presented in **Chapter 3** supports this idea. Several negative interactions were observed and it suggests that *S. aureus* may competitively exclude commensal species such as *S. epidermidis*. An alternative landscape of disturbances was observed in PSO with several species such as *Corynebacterium simulans*, *Peptostreptococcus anaerobius* and *Neisseriaceae G. sp.* displaying greater abundance on psoriatic skin, thus, it may be the case that multiple species operating in concert contribute towards the psoriatic phenotype.

Despite differences in pathogenic species, some general concepts of skin inflammation can be determined. These include a loss of potentially beneficial taxa such as *Propionibacterium* which were reduced on diseased skin. This species often colonises sebaceous sites [38] and it could be related to the the dryness associated with inflamed skin. Other possible homeostatic candidates included *Lactobacillus* which were lost in disease, although further work will need to be performed to determine its disease association with respect to gender which clearly impacts upon the abundance of this species. Taxa such as *Propionibacterium* were also reduced on non-lesional skin which suggests that potential probiotic treatments could help to restore species diversity or be used as a possible preventative treatment combating the onset of dysbiosis.

In comparison to healthy skin, the similarity between lesional and uninvolved skin was high. The community composition was only marginally different between ADNL and ADL, and no significant differences could be detected between PSONL and PSOL. A simple yet naive explanation could be that the composition of disease susceptible skin is already in a state of dysbiosis before an inflammatory triggering event. The data presented in this thesis does not allow us to make this conclusion as it is possible that, either during, or preceding a flare, the community composition may undergo a systemic shift such that the global cutaneous microbiome is altered during a flare phase. Furthermore, unlike the microbiota in which uninvolved sites closely resemble lesional skin, the transcriptomics analysis presented in **Chapter 4**, showed that uninvolved transcriptional profiles were much closer

to healthy skin. This observed 'lag' between the microbiota and transcriptome supports the idea that changes in the microbiome precede a flare and microbial dysbiosis could be a required precursor to inflammation. Sampling of patients in a flare-free phase could be performed to investigate this possibility, and if shifts in community composition occur during a pre-flare phase, longitudinal analyses may be able to determine when this occurs allowing for a causative analysis.

The findings in **Chapter 4** defined the transcriptional architecture of disease and established gene sets for integration with the cutaneous microbiota. The results demonstrated broad transcriptional changes in the lesional skin of both diseases. This analysis indicated that AD was characteristic of heightened Th2 activity, with disruptions to the epidermal barrier and extracellular matrix. Alternatively, PSO was associated with dysregulation of IL36, Th1/Th17 cytokines, antimicrobial peptides as well as disruptions to energy metabolism and axonal guidance signalling pathways. A considerable proportion of the inflammatory gene signatures overlapped which could be further investigated as therapeutic targets potentially effective for both conditions.

The matched microbiome and transcriptome data generated by the MAARS consortium provide a unique opportunity to study interactions between the host and resident microbiota, however, integration of high dimensional datasets is challenging. To achieve appropriate levels of sensitivity, powerful dimensionality reduction methods must be applied [29]. The overwhelming host-microbiota signal identified was between *S. aureus* and transcriptional profiles in AD. In **Chapters 5** and **6**, it was demonstrated that the impact of most bacteria on the host transcritpome was low, and only significant associations were identified between *S. aureus* and host transcripts in AD. These analyses point towards an intimate relationship that exists between *S. aureus* and the pathogenesis of atopic dermatitis. Negative associations were identified between *S. aureus* and IL34 which is critical for the development of Langerhans cells and has recently been associated with AD [211]. A cluster of genes encoding integral components of the epidermal barrier such as FLG, FLG2 and LOR were also negatively associated with *S. aureus* establishing a relationship between barrier dysregulation and pathogen colonisation. Studies have indicated that TEWL is significantly higher in AD patients colonised by *S. aureus* [206], and this analysis describes a set of genes which may be involved in this process.

Several positive associations with the immune system were identified which support a pro-inflammatory role for *S. aureus* in AD. Relationships were identified with the complement system, the expression of antimicrobial peptides and a gene cluster of which IL4R was the top hub. These findings indicate that *S. aureus* is closely intertwined with host immune activation and the Th2 response which drives AD. As well as associations with the immune system, one of the strongest *S. aureus* associations involved genes encoding components of the extracellular matrix. Fibrosis is a characteristic symptom of AD [208] and further work could be performed to determine how *S. aureus* is associated with this response, or if ECM remodelling is mostly due to increased immune activation [198]. Further associations were identified within a module encoding processes for non-coding RNA. There is limited literature regarding the role of ncRNAs in AD, therefore, this finding could represent a component of atopic inflammation that may have been overlooked and could be clarified through further research.

S. aureus is highly abundant on non-lesional skin, and accounts for approximately 17% of the uninvolved microbiota. Despite clear host-microbe associations in the lesional phase, no significant associations were identified on non-lesional skin. The reasons for this are unclear, however, it could relate to the observation that a lower abundance of staphylococcal enterotoxin is isolated from *S. aureus* strains on non-lesional skin, than from *S. aureus* strains on lesional skin [66]. Further analysis with metagenomics sequencing will be able to interrogate *S. aureus* at higher resolutions to the strain level enabling the investigation of this hypothesis.

Psoriasis is associated with mutations of the innate and adaptive immune system [73]. Mutations within NFκB signalling are also a factor of Crohn's disease and it is thought that enteric inflammation may be driven by inappropriate immune responses to intestinal microbiota [74, 80]. With the extreme transcriptional disturbances indicated in **Chapter 4** to the immune system, epidermal barrier and AMPs, it was expected that this dysregulation would be detectable in the relative proportions of the commensal microbiota. Instead, little covariation between the psoriatic microbiota and host gene expression profiles was observed and may represent an inability of the microbiome to shape host gene expression in the short term. Despite mostly inconclusive results for host-microbe associations in PSO, it cannot be concluded that the microbiome is unrelated to transcriptional profiles in the skin. *S.aureus* completely dominated the microbiota in AD and despite an average

relative abundance of $> 40\%$, the effect sizes observed for this pathogen were still fairly modest with maximum correlations of less then 0.5. When compared to *C. simulans* which accounted for only approximately 3% of the lesional psoriatic microbiota, it is reasonable to assume that effect size of this species amongst other potential pathogens in PSO would be considerably lower than *S. aureus*. I expect that more samples would be required to accurately elucidate host-microbe interactions in systems such as psoriasis where the community composition is only moderately disturbed.

It is important to note that 16S sequencing only profiles bacterial abundances and the cutaneous microbiome is also rich with fungal and viral microbiota [33]. *Malassezia* is one of the most prevalent fungi species on the skin [33] and has been linked to skin inflammation [206]. Further analysis could be performed to investigate the viral and fungal components which may be relevant to dysbiosis in skin inflammation.

Overall the results presented in this thesis contribute towards a greater understanding into the potential mechanisms which underlie dysbiosis in skin disease. The methods applied for integration of omics' datasets enabled the investigation into the relationships between pathogens and host parameters and provided insight into how these interactions may relate to inflammatory skin pathologies. Only with further validation and mechanistic studies will it be possible to translate these findings into clinical applications.

Bibliography

[1] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[2] Ann M O'Hara and Fergus Shanahan. The gut flora as a forgotten organ. *EMBO Reports*, 7(7):688–693, July 2006.

[3] Ron Sender, Shai Fuchs, and Ron Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8):e1002533, August 2016.

[4] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, February 2006.

[5] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012.

[6] Dirk Gevers, Subra Kugathasan, Lee A. Denson, et al. The Treatment-Naive Microbiome in New-Onset Crohns Disease. *Cell Host & Microbe*, 15(3):382–392, March 2014.

[7] Eric S. Lander, Lauren M. Linton, Bruce Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[8] C R Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–271, June 1987.

[9] P. C. Y. Woo, S. K. P. Lau, J. L. L. Teng, H Tse, and K. Y. Yuen. Then and now: use of 16s rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection*, 14(10):908–934, October 2008.

[10] F. Meyer, D. Paarmann, M. D'Souza, et al. The metagenomics RAST server a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[11] Qunfeng Dong, Jennifer M. Brulc, Alfonso Iovieno, et al. Diversity of Bacteria at Healthy Human Conjunctiva. *Investigative Ophthalmology & Visual Science*, 52(8):5408–5413, July 2011.

[12] J. Amar, M. Serino, C. Lange, et al. Involvement of tissue bacteria in the onset of diabetes in humans: evidence for a concept. *Diabetologia*, 54(12):3055–3061, December 2011.

[13] James M. Beck, Vincent B. Young, and Gary B. Huffnagle. The Microbiome of the Lung. *Translational research : the journal of laboratory and clinical medicine*, 160(4):258–266, October 2012.

[14] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, May 2011.

[15] Junjie Qin, Ruiqiang Li, Jeroen Raes, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010.

[16] Maria G. Dominguez-Bello, Elizabeth K. Costello, Monica Contreras, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America*, 107(26):11971–11975, June 2010.

[17] Julia K. Goodrich, Jillian L. Waters, Angela C. Poole, et al. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, November 2014.

[18] Brian D. Muegge, Justin Kuczynski, Dan Knights, et al. Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science*, 332(6032):970–974, May 2011.

[19] Jeremy E. Koenig, Aym Spor, Nicholas Scalfone, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl 1):4578–4585, March 2011.

[20] Carmen Haro, Oriol A. Rangel-Ziga, Juan F. Alcal-Daz, et al. Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PLOS ONE*, 11(5):e0154090, May 2016.

[21] Lora V. Hooper, Dan R. Littman, and Andrew J. Macpherson. Interactions between the microbiota and the immune system. *Science (New York, N.Y.)*, 336(6086):1268–1273, June 2012.

[22] June L. Round and Sarkis K. Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323, May 2009.

[23] Yun Kyung Lee and Sarkis K. Mazmanian. Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science (New York, N.Y.)*, 330(6012):1768–1773, December 2010.

[24] Jean-Franois Bach. The Effect of Infections on Susceptibility to Autoimmune and Allergic Diseases. *New England Journal of Medicine*, 347(12):911–920, September 2002.

[25] D. P. Strachan. Hay fever, hygiene, and household size. *BMJ : British Medical Journal*, 299(6710):1259–1260, November 1989.

[26] Les Dethlefsen, Sue Huse, Mitchell L. Sogin, and David A. Relman. The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16s rRNA Sequencing. *PLOS Biology*, 6(11):e280, November 2008.

[27] H Okada, C Kuhn, H Feillet, and J-F Bach. The hygiene hypothesis for autoimmune and allergic diseases: an update. *Clinical and Experimental Immunology*, 160(1):1–9, April 2010.

[28] Floris Imhann, Arnau Vich Vila, Marc Jan Bonder, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut*, pages gutjnl–2016–312135, October 2016.

[29] Xochitl C. Morgan, Boyko Kabakchiev, Levi Waldron, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*, 16:67, 2015.

[30] Frederic H. Martini. The Integumentary System: An Overview. In *Fundamentals of Anatomy & Physiology*, volume 7, pages 154–164. Pearson, 2006.

[31] Elizabeth A. Grice and Julia A. Segre. The skin microbiome. *Nature Reviews. Microbiology*, 9(4):244–253, April 2011.

[32] A. L. Cogen, V. Nizet, and R. L. Gallo. Skin microbiota: a source of disease or defence? *The British journal of dermatology*, 158(3):442–455, March 2008.

[33] Julia Oh, Allyson L. Byrd, Clay Deming, et al. Biogeography and individuality shape function in the human skin metagenome. *Nature*, 514(7520):59–64, October 2014.

[34] K. Wilke, A. Martin, L. Terstegen, and S. S. Biel. A short history of sweat gland biology. *International Journal of Cosmetic Science*, 29(3):169–179, June 2007.

[35] F. Niyonsaba, A. Suzuki, H. Ushio, et al. The human antimicrobial peptide dermcidin activates normal human keratinocytes. *British Journal of Dermatology*, 160(2):243–249, February 2009.

[36] Evgenia Makrantonaki, Ruta Ganceviciene, and Christos Zouboulis. An update on the role of the sebaceous gland in the pathogenesis of acne. *Dermato-endocrinology*, 3(1):41–49, 2011.

[37] Georg T. Wondrak. *Skin Stress Response Pathways: Environmental Factors and Molecular Opportunities*. Springer, August 2016. Google-Books-ID: ViXk-DAAAQBAJ.

[38] Elizabeth A. Grice, Heidi H. Kong, Sean Conlan, et al. Topographical and Temporal Diversity of the Human Skin Microbiome. *Science*, 324(5931):1190–1192, May 2009.

[39] Elizabeth K. Costello, Christian L. Lauber, Micah Hamady, et al. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science (New York, N.Y.)*, 326(5960):1694–1697, December 2009.

[40] Shi Ying, Dan-Ning Zeng, Liang Chi, et al. The Influence of Age and Gender on Skin-Associated Microbial Communities in Urban and Rural Human Populations. *PLOS ONE*, 10(10):e0141842, October 2015.

[41] Noah Fierer, Micah Hamady, Christian L. Lauber, and Rob Knight. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences*, 105(46):17994–17999, November 2008.

[42] Noah Fierer, Christian L. Lauber, Nick Zhou, et al. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6477–6481, April 2010.

[43] Shruti Naik, Nicolas Bouladoux, Christoph Wilhelm, et al. Compartmentalized Control of Skin Immunity by Resident Commensals. *Science (New York, N.Y.)*, 337(6098):1115–1119, August 2012.

[44] Thomas Bieber. Atopic Dermatitis. *Annals of Dermatology*, 22(2):125–137, May 2010.

[45] Anita Remitz and Sakari Reitamo. The clinical manifestations of atopic dermatitis. *Atopic textbook of*, page 1, 2008.

[46] F. V. Schultz Larsen and N. V. Holm. Atopic dermatitis in a population based twin series. Concordance rates and heritability estimation. *Acta Dermato-Venereologica. Supplementum*, 114:159, 1985.

[47] Jonathan M Spergel and Amy S Paller. Atopic dermatitis and the atopic march. *Journal of Allergy and Clinical Immunology*, 112(6, Supplement):S118–S127, December 2003.

[48] Donald Y.M. Leung, Mark Boguniewicz, Michael D. Howell, Ichiro Nomura, and Qutayba A. Hamid. New insights into atopic dermatitis. *Journal of Clinical Investigation*, 113(5):651–657, March 2004.

[49] Colin N. A. Palmer, Alan D. Irvine, Ana Terron-Kwiatkowski, et al. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nature Genetics*, 38(4):441–446, April 2006.

[50] Grainne M. O'Regan, Aileen Sandilands, W. H. Irwin McLean, and Alan D. Irvine. Filaggrin in atopic dermatitis. *The Journal of Allergy and Clinical Immunology*, 122(4):689–693, October 2008.

[51] Mark Boguniewicz and Donald YM Leung. Atopic Dermatitis: A Disease of Altered Skin Barrier and Immune Dysregulation. *Immunological reviews*, 242(1):233–246, July 2011.

[52] Padraic G. Fallon, Takashi Sasaki, Aileen Sandilands, et al. A homozygous frameshift mutation in the mouse Flg gene facilitates enhanced percutaneous allergen priming. *Nature Genetics*, 41(5):602–608, May 2009.

[53] Chung-Ching Chu, Paola Di Meglio, and Frank O. Nestle. Harnessing dendritic cells in inflammatory skin diseases. *Seminars in Immunology*, 23(1):28–41, February 2011.

[54] Vassili Soumelis, Pedro A. Reche, Holger Kanzler, et al. Human epithelial cells trigger dendritic cellmediated allergic inflammation by producing TSLP. *Nature Immunology*, 3(7):673–680, July 2002.

[55] Susanne Ebner, Van Anh Nguyen, Markus Forstner, et al. Thymic stromal lymphopoietin converts human epidermal Langerhans cells into antigen-presenting cells that induce proallergic T cells. *Journal of Allergy and Clinical Immunology*, 119(4):982–990, April 2007.

[56] Tilo Biedermann, Yuliya Skabytska, Susanne Kaesler, and Thomas Volz. Regulation of T Cell Immunity in Atopic Dermatitis by Microbes: The Yin and Yang of Cutaneous Inflammation. *Frontiers in Immunology*, 6, July 2015.

[57] W. Peng and N. Novak. Pathogenesis of atopic dermatitis. *Clinical & Experimental Allergy*, 45(3):566–574, March 2015.

[58] David A. Ewald, Dana Malajian, James G. Krueger, et al. Meta-analysis derived atopic dermatitis (MADAD) transcriptome defines a robust AD signature highlighting the involvement of atherosclerosis and lipid metabolism pathways. *BMC Medical Genomics*, 8, October 2015.

[59] G. S. Pilgram, D. C. Vissers, H. van der Meulen, et al. Aberrant lipid organization in stratum corneum of patients with atopic dermatitis and lamellar ichthyosis. *The Journal of Investigative Dermatology*, 117(3):710–717, September 2001.

[60] G. Imokawa, A. Abe, K. Jin, et al. Decreased level of ceramides in stratum corneum of atopic dermatitis: an etiologic factor in atopic dry skin? *The Journal of Investigative Dermatology*, 96(4):523–526, April 1991.

[61] M Surez-Farias, S Tintle, A Shemer, et al. Non-lesional atopic dermatitis (AD) skin is characterized by broad terminal differentiation defects and variable immune abnormalities. *The Journal of allergy and clinical immunology*, 127(4):954–64.e1–4, April 2011.

[62] J. J. Leyden, R. R. Marples, and A. M. Kligman. Staphylococcus aureus in the lesions of atopic dermatitis. *The British Journal of Dermatology*, 90(5):525–530, May 1974.

[63] J.q. Gong, L. Lin, T. Lin, et al. Skin colonization by Staphylococcus aureus in patients with eczema and atopic dermatitis and relevant combined topical therapy: a double-blind multicentre randomized controlled trial. *British Journal of Dermatology*, 155(4):680–687, October 2006.

[64] Heidi H. Kong, Julia Oh, Clay Deming, et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research*, 22(5):850–859, May 2012.

[65] Michael R. Williams and Richard L. Gallo. The Role of the Skin Microbiome in Atopic Dermatitis. *Current Allergy and Asthma Reports*, 15(11):65, November 2015.

[66] Zollner, Wichelhaus, Hartung, et al. Colonization with superantigen-producing Staphylococcus aureus is associated with increased severity of atopic dermatitis. *Clinical & Experimental Allergy*, 30(7):994–1000, July 2000.

[67] Anna L. Cogen, Kenshi Yamasaki, Katheryn M. Sanchez, et al. Selective antimicrobial action is provided by phenol-soluble modulins derived from Staphylococcus epidermidis, a normal resident of the skin. *The Journal of Investigative Dermatology*, 130(1):192–200, January 2010.

[68] Yuping Lai, Anna L. Cogen, Katherine A. Radek, et al. Activation of TLR2 by a small molecule produced by Staphylococcus epidermidis increases antimicrobial defense against bacterial skin infections. *The Journal of Investigative Dermatology*, 130(9):2211–2221, September 2010.

[69] Muya Shu, Yanhan Wang, Jinghua Yu, et al. Fermentation of Propionibacterium acnes, a Commensal Bacterium in the Human Skin Microbiome, as Skin Probiotics against Methicillin-Resistant Staphylococcus aureus. *PLOS ONE*, 8(2):e55380, February 2013.

[70] Tara D. Rachakonda, Clayton W. Schupp, and April W. Armstrong. Psoriasis prevalence among adults in the United States. *Journal of the American Academy of Dermatology*, 70(3):512–516, March 2014.

[71] Mark G. Lebwohl, Arthur Kavanaugh, April W. Armstrong, and Abby S. Van Voorhees. US Perspectives in the Management of Psoriasis and Psoriatic Arthritis: Patient and Physician Results from the Population-Based Multinational Assessment of Psoriasis and Psoriatic Arthritis (MAPP) Survey. *American Journal of Clinical Dermatology*, 17:87–97, 2016.

[72] Joel M. Gelfand and Howa Yeung. Metabolic Syndrome in Patients with Psoriatic Disease. *The Journal of rheumatology. Supplement*, 89:24–28, July 2012.

[73] Frank O. Nestle, Daniel H. Kaplan, and Jonathan Barker. Psoriasis. *New England Journal of Medicine*, 361(5):496–509, July 2009.

[74] Elisha D.O. Roberson and Anne M. Bowcock. Psoriasis genetics: breaking the barrier. *Trends in genetics : TIG*, 26(9):415–423, September 2010.

[75] Michelle A. Lowes, Toyoko Kikuchi, Judilyn Fuentes-Duculan, et al. Psoriasis Vulgaris Lesions Contain Discrete Populations of Th1 and Th17 T Cells. *Journal of Investigative Dermatology*, 128(5):1207–1211, May 2008.

[76] Frank O. Nestle, Paola Di Meglio, Jian-Zhong Qin, and Brian J. Nickoloff. Skin immune sentinels in health and disease. *Nature Reviews Immunology*, 9(10):679–691, October 2009.

[77] Gys J. de Jongh, Patrick L. J. M. Zeeuwen, Martina Kucharekova, et al. High Expression Levels of Keratinocyte Antimicrobial Proteins in Psoriasis Compared with Atopic Dermatitis. *Journal of Investigative Dermatology*, 125(6):1163–1173, December 2005.

[78] Rajan P. Nair, Kristina Callis Duffin, Cynthia Helms, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics*, 41(2):199–204, February 2009.

[79] A. B. Kimball, K. A. Papp, Y. Wasfi, et al. Long-term efficacy of ustekinumab in patients with moderate-to-severe psoriasis treated for up to 5 years in the PHOENIX

1 study. *Journal of the European Academy of Dermatology and Venereology: JEADV*, 27(12):1535–1545, December 2013.

[80] L. Fry, B.s. Baker, A.v. Powles, A. Fahlen, and L. Engstrand. Is chronic plaque psoriasis triggered by microbiota in the skin? *British Journal of Dermatology*, 169(1):47–52, July 2013.

[81] Rafael de Cid, Eva Riveira-Munoz, Patrick L. J. M. Zeeuwen, et al. Deletion of the late cornified envelope LCE3b and LCE3c genes as a susceptibility factor for psoriasis. *Nature Genetics*, 41(2):211–215, February 2009.

[82] Heiwa Kanamori, Masatsugu Tanaka, Hiroshi Kawaguchi, et al. Resolution of psoriasis following allogeneic bone marrow transplantation for chronic myelogenous leukemia: Case report and review of the literature. *American Journal of Hematology*, 71(1):41–44, September 2002.

[83] X Li, J Li, L Wang, et al. Transmission of psoriasis by allogeneic bone marrow transplantation and blood transfusion. *Blood Cancer Journal*, 5(3):e288, March 2015.

[84] Zhan Gao, Chi-hong Tseng, Bruce E. Strober, Zhiheng Pei, and Martin J. Blaser. Substantial Alterations of the Cutaneous Bacterial Biota in Psoriatic Lesions. *PLOS ONE*, 3(7):e2719, July 2008.

[85] Alexander V. Alekseyenko, Guillermo I. Perez-Perez, Aieska De Souza, et al. Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome*, 1:31, 2013.

[86] Annika Fahln, Lars Engstrand, Barbara S. Baker, Anne Powles, and Lionel Fry. Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. *Archives of Dermatological Research*, 304(1):15–22, November 2011.

[87] Xochitl C. Morgan and Curtis Huttenhower. Chapter 12: Human Microbiome Analysis. *PLOS Computational Biology*, 8(12):e1002808, December 2012.

[88] Wolfgang R Streit and Ruth A Schmitz. Metagenomics the key to the uncultured microbes. *Current Opinion in Microbiology*, 7(5):492–498, October 2004.

[89] Jill E. Clarridge. Impact of 16s rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews*, 17(4):840–862, October 2004.

[90] Ramya Srinivasan, Ulas Karaoz, Marina Volegova, et al. Use of 16s rRNA Gene for Identification of a Broad Range of Clinically Relevant Bacterial Pathogens. *PLOS ONE*, 10(2):e0117617, February 2015.

[91] Soumitesh Chakravorty, Danica Helb, Michele Burday, Nancy Connell, and David Alland. A detailed analysis of 16s ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2):330–339, May 2007.

[92] T. Z. DeSantis, P. Hugenholtz, N. Larsen, et al. Greengenes, a Chimera-Checked 16s rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072, July 2006.

[93] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[94] Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, December 2009.

[95] J. Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010.

[96] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–1650, July 2009.

[97] Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput Biol*, 10(4):e1003531, April 2014.

[98] Sophie J Weiss, Zhenjiang Xu, Amnon Amir, et al. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. Technical report, PeerJ PrePrints, 2015.

[99] Marie-Agns Dillies, Andrea Rau, Julie Aubert, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, November 2013.

[100] Joseph N. Paulson, O. Colin Stine, Hctor Corrada Bravo, and Mihai Pop. Robust methods for differential abundance analysis in marker gene surveys. *Nature methods*, 10(12):1200–1202, December 2013.

[101] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

[102] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

[103] Juan Jovel, Jordan Patterson, Weiwei Wang, et al. Characterization of the Gut Microbiome Using 16s or Shotgun Metagenomics. *Frontiers in Microbiology*, 7, April 2016.

[104] Anne Chao. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.

[105] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, July 1948.

[106] Jose U. Scher, Carles Ubeda, Alejandro Artacho, et al. Decreased Bacterial Diversity Characterizes an Altered Gut Microbiota in Psoriatic Arthritis and Resembles Dysbiosis of Inflammatory Bowel Disease. *Arthritis & rheumatology (Hoboken, N.J.)*, 67(1):128–139, January 2015.

[107] Catherine Lozupone and Rob Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, December 2005.

[108] J. Roger Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325–349, February 1957.

[109] Alban Ramette. Multivariate analyses in microbial ecology. *Fems Microbiology Ecology*, 62(2):142–160, November 2007.

[110] Anna Maria Seekatz, Krishna Rao, Kavitha Santhosh, and Vincent Bensan Young. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent Clostridium difficile infection. *Genome Medicine*, 8:47, 2016.

[111] Nicola Segata, Jacques Izard, Levi Waldron, et al. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12:R60, 2011.

[112] James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology*, 5(4):e1000352, April 2009.

[113] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, September 2008.

[114] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[115] Joseph N. Paulson, O. Colin Stine, Hctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, December 2013.

[116] Scott Schwartz, Iddo Friedberg, Ivan V Ivanov, et al. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biology*, 13(4):r32, 2012.

[117] Martin J Blaser, Maria G Dominguez-Bello, Monica Contreras, et al. Distinct cutaneous bacterial assemblages in a sampling of South American Amerindians and US residents. *The ISME Journal*, 7(1):85–95, January 2013.

[118] T Tickle, L Waldron, and C Huttenhower. Multivariate association of microbial communities with rich metadata in high-dimensional studies.

[119] Yael Haberman, Timothy L. Tickle, Phillip J. Dexheimer, et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *The Journal of Clinical Investigation*, 124(8):3617–3633, August 2014.

[120] Marcos Prez-Losada, Eduardo Castro-Nallar, Matthew L. Bendall, Robert J. Freishtat, and Keith A. Crandall. Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma. *PloS One*, 10(6):e0131819, 2015.

[121] Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, April 2016.

[122] Christian Hoffmann, Serena Dollive, Stephanie Grunberg, et al. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS ONE*, 8(6), June 2013.

[123] Zhanshan (Sam) Ma, Qiong Guan, Chengxi Ye, et al. Network analysis suggests a potentially evil alliance of opportunistic pathogens inhibited by a cooperative network in human milk bacterial communities. *Scientific Reports*, 5:8275, February 2015.

[124] Jonathan Friedman and Eric J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLOS Comput Biol*, 8(9):e1002687, September 2012.

[125] Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, et al. Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Computational Biology*, 8(7):e1002606, July 2012.

[126] Dennise D. DalmaWeiszhausz, Janet Warrington, Eugene Y. Tanimoto, and C. Garrett Miyada. [1] The Affymetrix GeneChip Platform: An Overview. volume 410 of *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, pages 3–28. Academic Press, 2006.

[127] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lnne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology*, 24(1):22–30, February 2013.

[128] John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32 Suppl:496–501, December 2002.

[129] Anis H. Khimani, Abner M. Mhashilkar, Alvydas Mikulskis, et al. Housekeeping genes in cancer: normalization of array data. *BioTechniques*, 38(5):739–745, May 2005.

[130] Earl Hubbell, Wei-Min Liu, and Rui Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, December 2002.

[131] Stephen R Piccolo, Ying Sun, Joshua D Campbell, et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 100(6):337–344, December 2012.

[132] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, April 2003.

[133] David B. Allison, Xiangqin Cui, Grier P. Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65, January 2006.

[134] Richard A. Miller, Andrzej Galecki, and Robert J. Shmookler-Reis. Interpretation, Design, and Analysis of Gene Array Expression Experiments. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(2):B52–B57, February 2001.

[135] Seo Young Kim, Jae Won Lee, and In Suk Sohn. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. *Statistical Methods in Medical Research*, 15(1):3–20, February 2006.

[136] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001.

[137] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3, 2004.

[138] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[139] Marine Jeanmougin, Aurelien de Reynies, Laetitia Marisa, et al. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLOS ONE*, 5(9):e12336, September 2010.

[140] Suyan Tian, James G. Krueger, Katherine Li, et al. Meta-Analysis Derived (MAD) Transcriptome of Psoriasis Defines the Core Pathogenesis of Disease. *PLOS ONE*, 7(9):e44274, September 2012.

[141] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[142] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[143] Ingenuity Systems. http://www.ingenuity.com.

[144] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[145] David Croft, Gavin OKelly, Guanming Wu, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database issue):D691–D697, January 2011.

[146] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology*, 8(2):e1002375, February 2012.

[147] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[148] Markus Ringnr. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, March 2008.

[149] John Tomfohr, Jun Lu, and Thomas B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, 2005.

[150] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article17, 2005.

[151] Tova F. Fuller, Anatole Ghazalpour, Jason E. Aten, et al. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 18(6-7):463–472, July 2007.

[152] M. J. Rothe and J. M. Grant-Kels. Diagnostic criteria for atopic dermatitis. *Lancet (London, England)*, 348(9030):769–770, September 1996.

[153] Christopher Quince, Anders Lanzen, Russell J. Davenport, and Peter J. Turnbaugh. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12:38, 2011.

[154] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26(19):2460–2461, October 2010.

[155] Robert C. Edgar, Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, August 2011.

[156] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayQualityMetricsa bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, 2009.

[157] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affyanalysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[158] Zhan Gao, Chi-hong Tseng, Zhiheng Pei, and Martin J. Blaser. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2927–2932, February 2007.

[159] Stephanie K. Lathrop, Seth M. Bloom, Sindhuja M. Rao, et al. Peripheral education of the immune system by colonic commensal microbiota. *Nature*, 478(7368):250–254, October 2011.

[160] Elizabeth R. Mann, Kathryn M. Smith, David Bernardo, et al. Review: Skin and the Immune System. *Journal of Clinical & Experimental Dermatology Research*, January 2012.

[161] Alexander Salava and Antti Lauerma. Role of the skin microbiome in atopic dermatitis. *Clinical and Translational Allergy*, 4, October 2014.

[162] Philip Dixon. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14(6):927–930, December 2003.

[163] Paul J. McMurdie and Susan Holmes. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, 8(4):e61217, April 2013.

[164] Paul Vos, George Garrity, Dorothy Jones, et al. *Bergey's Manual of Systematic Bacteriology: Volume 3: The Firmicutes*, volume 3. Springer Science & Business Media, 2011.

[165] Miron B Kursa, Witold R Rudnicki, and others. *Feature selection with the Boruta package*. Journal, 2010.

[166] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[167] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, October 2005.

[168] Paul Shannon, Andrew Markiel, Owen Ozier, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, November 2003.

[169] David Hgg, Marie Eriksson, Anders Sundstrm, and Marcus Schmitt-Egenolf. The Higher Proportion of Men with Psoriasis Treated with Biologics May Be Explained by More Severe Disease in Men. *PLoS ONE*, 8(5), May 2013.

[170] R. Toumi, I. Soufli, H. Rafa, et al. Probiotic bacteria lactobacillus and bifidobacterium attenuate inflammation in dextran sulfate sodium-induced experimental colitis in mice. *International Journal of Immunopathology and Pharmacology*, 27(4):615–627, December 2014.

[171] Thomas Bieber. Atopic Dermatitis. *New England Journal of Medicine*, 358(14):1483–1494, April 2008.

[172] M. Olsson, A. Broberg, M. Jerns, et al. Increased expression of aquaporin 3 in atopic eczema. *Allergy*, 61(9):1132–1137, September 2006.

[173] Mayte Surez-Farias, Benjamin Ungar, Joel Correa da Rosa, et al. RNA sequencing atopic dermatitis transcriptome profiling provides insights into novel disease mechanisms with potential therapeutic implications. *The Journal of Allergy and Clinical Immunology*, 135(5):1218–1227, May 2015.

[174] Douglas A. Plager, Alexey A. Leontovich, Susan A. Henke, et al. Early cutaneous gene transcription changes in adult atopic dermatitis and potential clinical implications. *Experimental Dermatology*, 16(1):28–36, January 2007.

[175] Yihong Yao, Laura Richman, Chris Morehouse, et al. Type I Interferon: Potential Therapeutic Target for Psoriasis? *PLOS ONE*, 3(7):e2737, July 2008.

[176] Johann E. Gudjonsson, Jun Ding, Andrew Johnston, et al. Assessment of the Psoriatic Transcriptome in a Large Sample: Additional Regulated Genes and Comparisons with In Vitro Models. *The Journal of investigative dermatology*, 130(7):1829–1840, July 2010.

[177] Johann E. Gudjonsson, Jun Ding, Xing Li, et al. Global Gene Expression Analysis Reveals Evidence for Decreased Lipid Biosynthesis and Increased Innate Immunity in Uninvolved Psoriatic Skin. *Journal of Investigative Dermatology*, 129(12):2795–2804, December 2009.

[178] Xianghong Zhou, James G. Krueger, Ming-Chih J. Kao, et al. Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiological Genomics*, 13(1):69–78, March 2003.

[179] Maris Keermann, Sulev Kks, Ene Reimann, et al. Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36rn. *BMC Genomics*, 16(1), April 2015.

[180] Mayte Surez-Farias, Katherine Li, Judilyn Fuentes-Duculan, et al. Expanding the Psoriasis Disease Profile: Interrogation of the Skin and Serum of Patients with Moderate-to-Severe Psoriasis. *The Journal of Investigative Dermatology*, 132(11):2552–2564, November 2012.

[181] Joachim Reischl, Susanne Schwenke, Johanna M. Beekman, et al. Increased Expression of Wnt5a in Psoriatic Plaques. *Journal of Investigative Dermatology*, 127(1):163–169, January 2007.

[182] Joel M. Gelfand and Howa Yeung. Metabolic Syndrome in Patients with Psoriatic Disease. *The Journal of rheumatology. Supplement*, 89:24–28, July 2012.

[183] Ichiro Nomura, Bifeng Gao, Mark Boguniewicz, et al. Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. *The Journal of Allergy and Clinical Immunology*, 112(6):1195–1202, December 2003.

[184] Emma Guttman-Yassky, Michelle A. Lowes, Judilyn Fuentes-Duculan, et al. Low expression of the IL-23/Th17 pathway in atopic dermatitis compared to psoriasis. *Journal of Immunology (Baltimore, Md.: 1950)*, 181(10):7420–7427, November 2008.

[185] Maria Quaranta, Bettina Knapp, Natalie Garzorz, et al. Intraindividual genome expression analysis reveals a specific molecular signature of psoriasis and eczema. *Science Translational Medicine*, 6(244):244ra90, July 2014.

[186] Julia K. Gittler, Avner Shemer, Mayte Surez-Farias, et al. Progressive activation of Th2/Th22 cytokines and selective epidermal proteins characterizes acute and chronic atopic dermatitis. *The Journal of allergy and clinical immunology*, 130(6):1344–1354, December 2012.

[187] Angelo Massimiliano D'Erme, Dagmar Wilsmann-Theis, Julia Wagenpfeil, et al. IL-36 (IL-1f9) Is a Biomarker for Psoriasis Skin Lesions. *Journal of Investigative Dermatology*, 135(4):1025–1032, April 2015.

[188] Emma Guttman-Yassky, Mayte Surez-Farias, Andrea Chiricozzi, et al. Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis. *Journal of Allergy and Clinical Immunology*, 124(6):1235–1244.e58, December 2009.

[189] Edward Y. Chen, Christopher M. Tan, Yan Kou, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, 2013.

[190] JE Gudjonsson, A Johnston, SW Stoll, et al. Evidence for altered Wnt signaling in psoriatic skin. *The Journal of investigative dermatology*, 130(7):1849–1859, July 2010.

[191] Noriko Ando, Yuki Nakamura, Rui Aoki, et al. Circadian Gene Clock Regulates Psoriasis-Like Skin Inflammation in Mice. *The Journal of Investigative Dermatology*, 135(12):3001–3008, December 2015.

[192] S.-E. Chang, S.-S. Han, H.-J. Jung, and J.-H. Choi. Neuropeptides and their receptors in psoriatic skin in relation to pruritus. *The British Journal of Dermatology*, 156(6):1272–1277, June 2007.

[193] Junko Yamaguchi, Michiko Aihara, Yusuke Kobayashi, Takeshi Kambara, and Zenro Ikezawa. Quantitative analysis of nerve growth factor (NGF) in the atopic dermatitis and psoriasis horny layer and effect of treatment on NGF in atopic dermatitis. *Journal of Dermatological Science*, 53(1):48–54, January 2009.

[194] J.C. Szepietowski and A. Reich. Pruritus in psoriasis: An update. *European Journal of Pain*, 20(1):41–46, January 2016.

[195] M. Nakamura, M. Toyoda, and M. Morohashi. Pruritogenic mediators in psoriasis vulgaris: comparative evaluation of itch-associated cutaneous factors. *The British Journal of Dermatology*, 149(4):718–730, October 2003.

[196] B. S. Bochner, D. A. Klunk, S. A. Sterbinsky, R. L. Coffman, and R. P. Schleimer. IL-13 selectively induces vascular cell adhesion molecule-1 expression in human endothelial cells. *Journal of Immunology (Baltimore, Md.: 1950)*, 154(2):799–803, January 1995.

[197] C.-C. E. Lan, A.-H. Fang, P.-H. Wu, and C.-S. Wu. Tacrolimus abrogates TGF-1-induced type I collagen production in normal human fibroblasts through suppressing p38mapk signalling pathway: implications on treatment of chronic atopic dermatitis lesions. *Journal of the European Academy of Dermatology and Venereology: JEADV*, 28(2):204–215, February 2014.

[198] Chun Geun Lee, Robert J. Homer, Zhou Zhu, et al. Interleukin-13 Induces Tissue Fibrosis by Selectively Stimulating and Activating Transforming Growth Factor 1. *The Journal of Experimental Medicine*, 194(6):809–822, September 2001.

[199] Luigi Tortola, Esther Rosenwald, Brian Abel, et al. Psoriasiform dermatitis is driven by IL-36-mediated DC-keratinocyte crosstalk. *The Journal of Clinical Investigation*, 122(11):3965–3976, November 2012.

[200] Risa Tamagawa-Mineoka, Yasutaro Okuzawa, Koji Masuda, and Norito Katoh. Increased serum levels of interleukin 33 in patients with atopic dermatitis. *Journal of the American Academy of Dermatology*, 70(5):882–888, May 2014.

[201] Natalie Garzorz, Linda Krause, Felix Lauffer, et al. A novel molecular disease classifier for psoriasis and eczema. *Experimental Dermatology*, May 2016.

[202] Hideki Kizawa, Ikuyo Kou, Aritoshi Iida, et al. An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nature Genetics*, 37(2):138–144, February 2005.

[203] Vitam Kodelja, Carola Mller, Oliver Politz, et al. Alternative Macrophage Activation-Associated CC-Chemokine-1, a Novel Structural Homologue of Macrophage Inflammatory Protein-1 with a Th2-Associated Expression Pattern. *The Journal of Immunology*, 160(3):1411–1418, February 1998.

[204] Takashi Yoshiike, Yosuke Aikawa, Jirot Sindhvananda, et al. Skin barrier defect in atopic dermatitis: increased permeability of the stratum corneum using dimethyl sulfoxide and theophylline. *Journal of Dermatological Science*, 5(2):92–96, April 1993.

[205] Mark Boguniewicz and Donald YM Leung. Atopic Dermatitis: A Disease of Altered Skin Barrier and Immune Dysregulation. *Immunological reviews*, 242(1):233–246, July 2011.

[206] Camilla Ling Jinnestl, Emma Belfrage, Ove Bck, Artur Schmidtchen, and Andreas Sonesson. Skin barrier impairment correlates with cutaneous Staphylococcus aureus colonization and sensitization to skin-associated microbial antigens in adult patients with atopic dermatitis. *International Journal of Dermatology*, 53(1):27–33, January 2014.

[207] Michael P. Schn and W.-Henning Boehncke. Psoriasis. *New England Journal of Medicine*, 352(18):1899–1912, May 2005.

[208] Min-Hee Oh, Sun Young Oh, Jinho Yu, et al. IL-13 Induces Skin Fibrosis in Atopic Dermatitis by Thymic Stromal Lymphopoietin. *Journal of Immunology (Baltimore, Md. : 1950)*, 186(12):7232–7242, June 2011.

[209] Lorena Riol-Blanco, Jose Ordovas-Montanes, Mario Perro, et al. Nociceptive sensory neurons drive interleukin-23-mediated psoriasiform skin inflammation. *Nature*, 510(7503):157–161, June 2014.

[210] Riikka Kivel, Mika Silvennoinen, Maarit Lehti, et al. Gene expression centroids that link with low intrinsic aerobic exercise capacity and complex disease risk. *The FASEB Journal*, 24(11):4565–4574, November 2010.

[211] Hitokazu Esaki, David A. Ewald, Benjamin Ungar, et al. Identification of Novel Immune and Barrier Genes in Atopic Dermatitis by Laser Capture Micro-dissection. *The Journal of allergy and clinical immunology*, 135(1):153–163, January 2015.

[212] Allison-Lynn Andrews, John W. Holloway, Stephen T. Holgate, and Donna E. Davies. IL-4 receptor alpha is an important modulator of IL-4 and IL-13 receptor binding: implications for the development of therapeutic targets. *Journal of Immunology (Baltimore, Md.: 1950)*, 176(12):7456–7461, June 2006.

[213] K. M. Cunnion and M. M. Frank. Complement Activation Influences Staphylococcus aureus Adherence to Endothelial Cells. *Infection and Immunity*, 71(3):1321–1327, March 2003.

[214] Lydia Sorokin. The impact of the extracellular matrix on inflammation. *Nature Reviews Immunology*, 10(10):712–723, October 2010.

[215] Nevena Skroza, Ilaria Proietti, Riccardo Pampena, et al. Correlations between Psoriasis and Inflammatory Bowel Diseases. *BioMed Research International*, 2013:e983902, July 2013.

[216] Melanie Greter, Iva Lelios, Pawel Pelczar, et al. Stroma-derived interleukin-34 controls the development and maintenance of langerhans cells and the maintenance of microglia. *Immunity*, 37(6):1050–1060, December 2012.

[217] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, December 1998.

[218] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, June 2002.

[219] Adriano V. Werhli, Marco Grzegorczyk, and Dirk Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, October 2006.

[220] Adam A Margolin, Ilya Nemenman, Katia Basso, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, March 2006.

[221] Julien Chiquet, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25(3):417–418, February 2009.

[222] Jeffrey D. Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing Statistical Methods for Constructing Large Scale Gene Networks. *PLOS ONE*, 7(1):e29348, January 2012.

[223] Jeremy A. Miller, Song-Lin Ding, Susan M. Sunkin, et al. Transcriptional landscape of the prenatal human brain. *Nature*, 508(7495):199–206, April 2014.

[224] Michael C. Oldham, Steve Horvath, and Daniel H. Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47):17973–17978, November 2006.

[225] Andreas Neueder and Gillian P. Bates. A common gene expression signature in Huntingtons disease patient brain regions. *BMC Medical Genomics*, 7:60, 2014.

[226] Irina Voineagu, Xinchen Wang, Patrick Johnston, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384, June 2011.

[227] Frederick E. Dewey, Marco V. Perez, Matthew T. Wheeler, et al. Gene Coexpression Network Topology of Cardiac Development, Hypertrophy, and Failure. *Circulation. Cardiovascular Genetics*, 4(1):26–35, February 2011.

[228] Danning He, Zhi-Ping Liu, Masao Honda, Shuichi Kaneko, and Luonan Chen. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. *Journal of Molecular Cell Biology*, 4(3):140–152, June 2012.

[229] Akshata R. Udyavar, Megan D. Hoeksema, Jonathan E. Clark, et al. Co-expression network analysis identifies Spleen Tyrosine Kinase (SYK) as a candidate oncogenic driver in a subset of small-cell lung cancer. *BMC Systems Biology*, 7(5):1–16, 2013.

[230] Mamata F Khirade, Girdhari Lal, and Sharmila A Bapat. Derivation of a fifteen gene prognostic panel for six cancers. *Scientific reports*, 5, 2015.

[231] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[232] Peter Langfelder, Rui Luo, Michael C. Oldham, and Steve Horvath. Is My Network Module Preserved and Reproducible? *PLOS Comput Biol*, 7(1):e1001057, January 2011.

[233] Johanna Hardin, Aya Mitani, Leanne Hicks, and Brian VanKoten. A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, 8:220, 2007.

[234] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabsi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[235] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, March 2008.

[236] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, August 2000.

[237] Jun Dong and Steve Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1:24, 2007.

[238] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology*, 1:54, 2007.

[239] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, 16(5):284–287, May 2012.

[240] Atanas Kamburov, Ulrich Stelzl, Hans Lehrach, and Ralf Herwig. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue):D793–800, January 2013.

[241] Michael C. Oldham, Genevieve Konopka, Kazuya Iwamoto, et al. Functional organization of the transcriptome in human brain. *Nature Neuroscience*, 11(11):1271–1282, November 2008.

[242] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[243] Richard Ahn, Rashmi Gupta, Kevin Lai, et al. Network analysis of psoriasis reveals biological pathways and roles for coding and long non-coding RNAs. *BMC Genomics*, 17:841, 2016.

[244] Margaret Alexander and Ryan M. O'Connell. Noncoding RNAs and chronic inflammation: Micro-managing the fire within. *BioEssays*, 37(9):1005–1015, September 2015.

[245] Monika Krampert, Sandra Kuenzle, Shelley N.-M. Thai, et al. ADAMTS1 proteinase is up-regulated in wounded skin and regulates migration of fibroblasts and endothelial cells. *The Journal of Biological Chemistry*, 280(25):23844–23852, June 2005.

[246] Bon-Hun Koo, David M. Coe, Laura J. Dixon, et al. ADAMTS9 Is a Cell-Autonomously Acting, Anti-Angiogenic Metalloprotease Expressed by Microvascular Endothelial Cells. *The American Journal of Pathology*, 176(3):1494–1504, March 2010.

[247] Siva Kanangat, Arnold Postlethwaite, Karen Hasty, et al. Induction of multiple matrix metalloproteinases in human dermal and synovial fibroblasts by Staphylococcus aureus: implications in the pathogenesis of septic arthritis and other soft tissue infections. *Arthritis Research & Therapy*, 8(6):R176, 2006.

[248] Thomas A Wynn and Thirumalai R Ramalingam. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nature Medicine*, 18(7):1028–1040, July 2012.

[249] K. Matsui and A. Nishikawa. Peptidoglycan from Staphylococcus aureus induces T(H)2 immune response in mice. *Journal of Investigational Allergology & Clinical Immunology*, 22(2):80–86, 2012.

Appendix A

Supplementary information for Chapter 3

This appendix contains supplemental data for the independent microbiome analysis.

Group	Gender (padj)	Institution (padj)	Bodysite (padj)	Age (padj)	OTU
CTRL	9.93E-06	ns	ns	ns	Anaerococcus sp.
CTRL	5.46E-05	ns	ns	ns	Corynebacterium sp.
CTRL	2.38E-03	ns	ns	ns	Lactobacillus sp.
CTRL	5.67E-06	1.11E-02	ns	1.11E-02	Lactobacillus iners
CTRL	2.58E-02	ns	ns	ns	Prevotella sp.
CTRL	4.56E-03	4.15E-03	ns	ns	Lactobacillus sp.
CTRL	ns	5.04E-10	ns	5.66E-05	Corynebacterium sp.
CTRL	ns	1.66E-03	ns	ns	Bradyrhizobium sp.
CTRL	ns	1.89E-03	ns	ns	Anaerococcus sp.
CTRL	ns	1.27E-02	3.70E-10	ns	Herbaspirillum sp.
CTRL	ns	6.58E-06	ns	ns	Paracoccus marcusii
CTRL	ns	4.51E-03	ns	ns	Anaerococcus sp.
CTRL	ns	7.46E-03	5.95E-05	ns	Variovorax paradoxus
CTRL	ns	7.90E-08	ns	ns	Acinetobacter sp.
CTRL	ns	4.10E-03	ns	1.09E-03	Comamonadaceae G. sp.
CTRL	ns	1.01E-03	6.53E-05	ns	Phyllobacterium sp.
CTRL	ns	2.73E-03	ns	ns	Corynebacterium sp.
CTRL	ns	8.68E-06	ns	ns	Dermacoccus sp.
CTRL	ns	9.28E-03	3.16E-03	ns	Finegoldia sp.
CTRL	ns	3.63E-03	ns	ns	Micrococcus sp.
CTRL	ns	4.80E-04	ns	ns	Prevotella sp.
CTRL	ns	4.08E-03	ns	ns	Acinetobacter johnsonii
CTRL	ns	2.70E-03	ns	ns	Janibacter sp.
CTRL	ns	3.89E-12	ns	ns	Methylophilaceae G. sp.
CTRL	ns	1.86E-05	ns	9.33E-04	Dietzia sp.
CTRL	ns	2.62E-04	ns	ns	Acinetobacter sp.
CTRL	ns	ns	7.49E-04	ns	Staphylococcus aureus
CTRL	ns	ns	ns	2.67E-02	Corynebacterium sp.
AD	ns	1.56E-02	ns	ns	Acinetobacter sp.
AD	ns	1.04E-03	ns	ns	Corynebacterium sp.
AD	ns	7.93E-04	ns	ns	Acinetobacter sp.
AD	ns	2.79E-03	ns	ns	Lactococcus sp.
AD	ns	1.26E-02	ns	ns	Micrococcus sp.
AD	ns	1.59E-02	ns	ns	Streptococcus sp.
AD	ns	3.51E-02	ns	ns	Corynebacterium sp.

AD	ns	5.79E-05	ns	ns	Dietzia sp.
AD	ns	ns	4.41E-02	ns	Variovorax paradoxus
PSO	2.06E-02	ns	ns	ns	Corynebacterium sp.
PSO	ns	7.10E-04	ns	ns	Corynebacterium sp.
PSO	ns	2.87E-07	1.89E-02	ns	Acinetobacter sp.
PSO	ns	2.44E-04	ns	ns	Acinetobacter johnsonii
PSO	ns	5.07E-05	ns	ns	Dietzia sp.
PSO	ns	ns	3.73E-02	ns	Phyllobacterium sp.
PSO	ns	ns	4.24E-02	ns	Kocuria palustris
PSO	ns	ns	ns	4.78E-02	Corynebacterium kroppenstedtii

Table A.1: OTU-metadata associations

Tax Level	Taxa	Wilcoxon (padj)	MaAsLin (padj)	Coef
Phylum	Firmicutes	4.00E-14	1.20E-22	5.10E-01
Phylum	Proteobacteria	1.10E-10	9.50E-13	-3.90E-01
Phylum	Actinobacteria	1.00E-05	2.90E-05	-1.60E-01
Phylum	Bacteroidetes	1.10E-05	3.70E-05	-6.70E-02
Phylum	Cyanobacteria	7.50E-03	2.50E-04	-2.30E-02
Class	Bacilli	1.20E-15	8.80E-28	5.70E-01
Class	Alphaproteobacteria	2.90E-09	3.90E-10	-1.50E-01
Class	Betaproteobacteria	3.70E-13	1.50E-08	-2.80E-01
Class	Actinobacteria	1.10E-05	2.90E-05	-1.60E-01
Class	Clostridia	9.80E-07	6.10E-05	-8.30E-02
Class	Deltaproteobacteria	4.90E-05	9.30E-05	-1.40E-02
Class	[Saprospirae]	3.30E-04	3.30E-04	-1.30E-02
Class	Cytophagia	4.00E-02	2.10E-03	-8.10E-03
Class	Gammaproteobacteria	1.00E-04	7.20E-03	-1.20E-01
Class	Flavobacteriia	3.60E-03	1.10E-02	-3.20E-02
Class	Bacteroidia	2.10E-03	1.70E-02	-2.70E-02
Order	Bacillales	3.90E-17	3.20E-34	6.20E-01
Order	Burkholderiales	2.20E-13	2.80E-08	-2.70E-01
Order	Rhizobiales	3.90E-08	3.70E-07	-1.20E-01
Order	Actinomycetales	1.30E-05	2.90E-05	-1.60E-01
Order	Caulobacterales	2.60E-08	6.50E-05	-2.60E-02
Order	Sphingomonadales	1.10E-06	7.10E-05	-2.60E-02
Order	Clostridiales	1.20E-06	8.10E-05	-8.10E-02
Order	[Saprospirales]	3.30E-04	3.30E-04	-1.30E-02
Order	Methylophilales	2.40E-04	5.60E-04	-1.50E-02
Order	Lactobacillales	1.80E-05	7.60E-04	-8.50E-02
Order	Cytophagales	4.00E-02	2.10E-03	-8.10E-03
Order	Xanthomonadales	4.40E-04	2.30E-03	-2.20E-02
Order	Rhodobacterales	2.80E-07	3.40E-03	-2.80E-02
Order	Flavobacteriales	3.60E-03	1.10E-02	-3.20E-02
Order	Bacteroidales	2.10E-03	1.70E-02	-2.70E-02

Order	Rhodospirillales	4.30E-02	2.40E-02	-8.80E-03
Order	Pseudomonadales	2.10E-06	2.70E-02	-7.30E-02
Order	Pasteurellales	2.10E-02	3.50E-02	-1.10E-02
Family	Staphylococcaceae	6.30E-17	3.20E-34	6.30E-01
Family	Burkholderiaceae	2.00E-05	1.60E-06	-6.80E-02
Family	Propionibacteriaceae	7.30E-09	1.80E-05	-2.50E-02
Family	Brucellaceae	3.30E-04	2.00E-05	-1.80E-02
Family	Comamonadaceae	6.10E-11	6.50E-05	-2.00E-01
Family	Caulobacteraceae	2.60E-08	6.50E-05	-2.60E-02
Family	Sphingomonadaceae	4.90E-06	1.70E-04	-2.50E-02
Family	[Tissierellaceae]	3.50E-07	2.20E-04	-6.70E-02
Family	Bradyrhizobiaceae	2.00E-09	3.30E-04	-4.10E-02
Family	Chitinophagaceae	3.20E-04	4.70E-04	-1.30E-02
Family	Methylophilaceae	2.40E-04	5.60E-04	-1.50E-02
Family	Micrococcaceae	3.30E-05	7.60E-04	-8.70E-02
Family	Oxalobacteraceae	2.00E-02	1.90E-03	-5.70E-02
Family	Cytophagaceae	4.30E-02	2.00E-03	-8.10E-03
Family	Nocardioidaceae	9.40E-05	2.10E-03	-9.30E-03
Family	Clostridiaceae	6.60E-03	2.70E-03	-1.70E-02
Family	Aerococcaceae	2.10E-04	3.00E-03	-2.20E-02
Family	Xanthomonadaceae	6.80E-04	3.40E-03	-2.10E-02
Family	Rhodobacteraceae	4.90E-07	6.00E-03	-2.60E-02
Family	Rhizobiaceae	5.10E-04	6.80E-03	-1.30E-02
Family	Lactobacillaceae	3.00E-07	7.50E-03	-4.70E-02
Family	Corynebacteriaceae	2.30E-03	8.10E-03	-7.30E-02
Family	Lachnospiraceae	8.70E-03	9.40E-03	-1.50E-02
Family	Ruminococcaceae	2.80E-03	1.40E-02	-1.40E-02
Family	[Weeksellaceae]	4.90E-03	1.50E-02	-2.90E-02
Family	Microbacteriaceae	2.70E-02	3.10E-02	-9.30E-03
Family	Pasteurellaceae	2.10E-02	3.50E-02	-1.10E-02
Family	Moraxellaceae	4.00E-06	3.50E-02	-6.90E-02
Family	Prevotellaceae	2.20E-02	3.50E-02	-1.90E-02
Genus	Staphylococcus	9.60E-17	3.20E-34	6.30E-01
Genus	Burkholderia	6.20E-06	1.30E-06	-6.90E-02
Genus	Ochrobactrum	3.30E-04	1.90E-05	-1.80E-02
Genus	Propionibacterium	7.30E-09	2.10E-05	-2.50E-02
Genus	Anaerococcus	1.70E-07	4.10E-05	-5.70E-02
Genus	Micrococcus	4.50E-05	1.20E-03	-8.10E-02
Genus	Tepidimonas	2.20E-03	1.30E-03	-1.70E-02
Genus	Finegoldia	1.10E-05	2.10E-03	-2.80E-02
Genus	Bradyrhizobium	4.90E-08	2.70E-03	-3.50E-02

Genus	Thermoanaerobacterium	4.90E-03	3.00E-03	-9.90E-03
Genus	Acinetobacter	1.40E-05	5.10E-03	-4.50E-02
Genus	Lactobacillus	2.60E-07	7.50E-03	-4.70E-02
Genus	Corynebacterium	2.30E-03	8.10E-03	-7.30E-02
Genus	Variovorax	4.80E-02	1.10E-02	-3.10E-02
Genus	Enhydrobacter	4.90E-05	1.10E-02	-5.40E-02
Genus	Pelomonas	3.60E-07	1.20E-02	-2.50E-02
Genus	Paracoccus	1.10E-05	1.40E-02	-2.40E-02
Genus	Sphingomonas	4.40E-04	1.50E-02	-1.20E-02
Genus	Facklamia	9.90E-04	1.60E-02	-1.20E-02
Genus	Kocuria	1.30E-03	1.90E-02	-2.10E-02
Genus	Prevotella	2.20E-02	3.50E-02	-1.90E-02
Otu	Staphylococcus aureus	1.10E-29	1.70E-35	6.90E-01
Otu	Burkholderia sp.	3.70E-09	7.60E-09	-7.60E-02
Otu	Staphylococcus sp.	3.00E-11	3.30E-07	-7.50E-02
Otu	Ochrobactrum sp.	7.50E-04	4.40E-05	-1.80E-02
Otu	Sphingomonadaceae G. sp.	2.80E-03	6.50E-05	-1.40E-02
Otu	Propionibacterium acnes	2.10E-07	9.40E-05	-2.30E-02
Otu	Methylophilaceae G. sp.	6.30E-04	5.90E-04	-1.50E-02
Otu	Anaerococcus sp.	3.10E-04	2.00E-03	-1.10E-02
Otu	Corynebacterium sp.	1.30E-03	2.00E-03	-2.70E-02
Otu	Tepidimonas sp.	2.20E-03	2.10E-03	-1.60E-02
Otu	Finegoldia sp.	1.10E-05	2.10E-03	-2.80E-02
Otu	Corynebacterium sp.	5.50E-03	3.70E-03	-1.90E-02
Otu	Thermoanaerobacterium sp.	2.60E-03	3.70E-03	-7.90E-03
Otu	Micrococcus sp.	9.40E-04	3.70E-03	-7.00E-02
Otu	Bradyrhizobium sp.	3.80E-07	5.10E-03	-3.30E-02
Otu	Anaerococcus sp.	9.80E-03	8.50E-03	-1.20E-02
Otu	Enhydrobacter sp.	6.50E-05	1.30E-02	-5.30E-02
Otu	Kocuria palustris	2.20E-05	1.40E-02	-1.30E-02
Otu	Anaerococcus sp.	1.40E-04	1.70E-02	-1.30E-02
Otu	Pelomonas sp.	1.50E-07	1.90E-02	-2.30E-02
Otu	Caulobacteraceae G. sp.	4.60E-03	2.70E-02	-1.30E-02
Otu	Comamonadaceae G. sp.	1.60E-04	4.40E-02	-1.00E-01

Table A.2: Significant ADL associated taxa

Tax Level	Taxa	Wilcoxon (padj)	MaAsLin (padj)	Coef
Phylum	Firmicutes	5.70E-04	2.30E-05	2.20E-01
Phylum	Proteobacteria	4.60E-03	3.10E-03	-1.90E-01
Class	Bacilli	1.00E-04	1.00E-06	2.40E-01

Class	Alphaproteobacteria	1.20E-02	1.50E-03	-8.30E-02
Class	Betaproteobacteria	7.60E-06	3.10E-03	-1.70E-01
Class	[Saprospirae]	1.20E-02	1.80E-02	-1.00E-02
Order	Bacillales	8.20E-08	1.60E-11	2.70E-01
Order	Rhizobiales	5.00E-03	1.50E-03	-8.00E-02
Order	Burkholderiales	1.00E-05	3.10E-03	-1.70E-01
Order	[Saprospirales]	1.20E-02	1.80E-02	-1.00E-02
Order	Methylophilales	8.10E-03	1.90E-02	-1.20E-02
Order	Lactobacillales	7.00E-04	2.70E-02	-5.80E-02
Family	Staphylococcaceae	8.20E-08	3.50E-12	2.90E-01
Family	Chitinophagaceae	1.20E-02	1.80E-02	-1.00E-02
Family	Methylophilaceae	8.10E-03	1.90E-02	-1.20E-02
Family	Comamonadaceae	6.50E-05	3.00E-02	-1.30E-01
Family	Propionibacteriaceae	6.50E-05	3.00E-02	-1.50E-02
Genus	Staphylococcus	8.20E-08	3.40E-12	2.90E-01
Genus	Burkholderia	3.10E-02	1.30E-02	-4.50E-02
Genus	Propionibacterium	9.20E-05	3.70E-02	-1.40E-02
Otu	Staphylococcus aureus	4.50E-22	2.40E-16	3.00E-01
Otu	Staphylococcus sp.	8.20E-08	3.90E-04	-6.50E-02
Otu	Burkholderia sp.	6.40E-05	5.70E-04	-5.30E-02
Otu	Corynebacterium sp.	7.00E-04	2.50E-02	1.10E-02
Otu	Methylophilaceae G. sp.	1.70E-02	3.00E-02	-1.10E-02

Table A.3: Significant ADNL associated taxa

Tax Level	Taxa	Wilcoxon (padj)	MaAsLin (padj)	Coef
Phylum	Firmicutes	5.60E-03	3.60E-02	1.00E-01
Phylum	Proteobacteria	7.70E-05	4.90E-02	-1.20E-01
Order	Neisseriales	1.40E-04	1.80E-02	1.40E-02
Order	Methylophilales	3.20E-04	4.40E-02	-1.00E-02
Family	Actinomycetaceae	9.10E-06	5.10E-04	2.40E-02
Family	Peptostreptococcaceae	9.30E-07	5.10E-04	2.00E-02
Family	Propionibacteriaceae	8.20E-03	5.10E-04	-2.10E-02
Family	Rhizobiaceae	4.30E-03	1.50E-02	-1.30E-02
Family	Neisseriaceae	1.40E-04	1.80E-02	1.40E-02
Family	Veillonellaceae	9.00E-04	2.20E-02	1.80E-02
Family	Prevotellaceae	4.90E-04	2.40E-02	2.70E-02
Family	Fusobacteriaceae	4.90E-04	2.80E-02	9.50E-03
Family	Corynebacteriaceae	2.50E-05	4.40E-02	6.80E-02
Family	Carnobacteriaceae	4.30E-03	4.40E-02	1.10E-02
Family	Methylophilaceae	3.20E-04	4.40E-02	-1.00E-02

Genus	Peptostreptococcus	4.20E-07	1.20E-07	1.90E-02
Genus	Actinomyces	9.30E-07	5.10E-04	2.10E-02
Genus	Propionibacterium	1.10E-02	6.30E-04	-2.00E-02
Genus	Porphyromonas	1.90E-03	2.20E-02	1.50E-02
Genus	Prevotella	4.90E-04	2.40E-02	2.70E-02
Genus	Peptoniphilus	2.30E-03	2.70E-02	2.60E-02
Genus	Fusobacterium	4.90E-04	2.80E-02	9.50E-03
Genus	Corynebacterium	2.50E-05	4.40E-02	6.80E-02
Genus	Streptococcus	2.30E-03	4.40E-02	3.10E-02
Otu	Corynebacterium simulans	5.70E-18	2.10E-09	3.50E-02
Otu	Peptostreptococcus anaerobius	1.50E-04	2.30E-05	1.50E-02
Otu	Neisseriaceae G. sp.	6.70E-08	5.80E-04	8.50E-03
Otu	Propionibacterium acnes	3.60E-02	1.10E-03	-2.00E-02
Otu	Streptococcus sp.	2.30E-07	6.10E-03	1.20E-02
Otu	Corynebacterium kroppenstedtii	2.20E-06	1.10E-02	1.80E-02
Otu	Prevotella sp.	5.60E-04	1.10E-02	8.90E-03
Otu	Prevotella sp.	6.80E-03	2.00E-02	1.10E-02
Otu	Corynebacterium kroppenstedtii	1.10E-06	2.20E-02	2.20E-02
Otu	Corynebacterium sp.	6.60E-05	2.40E-02	1.30E-02
Otu	Anaerococcus sp.	5.10E-03	2.50E-02	1.10E-02
Otu	Burkholderia sp.	1.40E-04	3.00E-02	-3.40E-02

Table A.4: Significant PSOL associated taxa

Tax Level	Taxa	Wilcoxon (padj)	MaAsLin (padj)	Coef
Order	Neisseriales	5.00E-04	3.80E-03	1.90E-02
Family	Actinomycetaceae	1.30E-03	2.40E-03	2.10E-02
Family	Neisseriaceae	5.00E-04	3.80E-03	1.90E-02
Family	Peptostreptococcaceae	2.40E-03	1.00E-02	1.50E-02
Family	Fusobacteriaceae	2.40E-03	4.80E-02	8.40E-03
Genus	Peptostreptococcus	1.30E-03	4.40E-04	1.30E-02
Genus	Actinomyces	3.30E-04	9.70E-04	2.00E-02
Genus	Porphyromonas	4.20E-02	4.30E-02	1.40E-02
Genus	Fusobacterium	2.40E-03	4.80E-02	8.40E-03
Otu	Corynebacterium simulans	6.40E-15	7.00E-10	3.70E-02
Otu	Neisseriaceae G. sp.	3.50E-08	4.40E-04	1.90E-02
Otu	Peptostreptococcus anaerobius	7.70E-03	1.40E-03	1.20E-02
Otu	Anaerococcus sp.	1.00E-03	2.10E-03	1.50E-02
Otu	Burkholderia sp.	2.10E-02	7.10E-03	-4.00E-02
Otu	Streptococcus sp.	3.30E-06	1.10E-02	1.10E-02
Otu	Corynebacterium kroppenstedtii	5.50E-05	1.30E-02	1.40E-02

Otu	Lactobacillus sp.	8.90E-09	4.80E-02	-1.10E-02
Otu	Corynebacterium sp.	6.00E-04	4.80E-02	1.20E-02

Table A.5: Significant PSONL associated taxa

Level	Taxa	Wilx P(U)	Mas P(U)	Coef(U)	Wilx P(M)	Mas P(M)	Coef(M)
Order	Pasteurellales	2.13E-02	3.47E-02	-1.11E-02	1.22E-01	3.35E-02	-1.14E-02
Family	Microbacteriaceae	2.73E-02	3.13E-02	-9.29E-03	3.61E-02	8.56E-02	-9.17E-03
Family	Pasteurellaceae	2.13E-02	3.47E-02	-1.11E-02	1.22E-01	3.35E-02	-1.14E-02
Family	Moraxellaceae	4.03E-06	3.47E-02	-6.93E-02	4.86E-05	5.70E-02	-5.66E-02
Family	Prevotellaceae	2.23E-02	3.47E-02	-1.90E-02	6.27E-02	1.74E-02	-1.96E-02
Family	Lachnospiraceae	8.71E-03	9.35E-03	-1.50E-02	5.92E-02	6.29E-03	-1.54E-02
Genus	Prevotella	2.23E-02	3.47E-02	-1.90E-02	6.27E-02	1.74E-02	-1.96E-02
Genus	Variovorax	4.78E-02	1.06E-02	-3.13E-02	1.33E-01	4.39E-03	-3.90E-02
Otu	Anaerococcus sp.	9.83E-03	8.47E-03	-1.24E-02	9.12E-02	2.26E-03	-1.18E-02
Otu	Caulobacteraceae G. sp.	4.65E-03	2.66E-02	-1.30E-02	5.03E-02	4.51E-02	-1.17E-02

Table A.6: Significant taxa for ADL-CTRL in the unmatched (U) but not significant in matched cohort (M). Wilx and Mas correspond to Wilcoxon ranked sum and MaAsLin respectively

Level	Taxa	Wilx P(U)	Mas P(U)	Coef(U)	Wilx P(M)	Mas P(M)	Coef(M)
Phylum	Firmicutes	5.59E-03	3.62E-02	1.01E-01	3.40E-02	5.97E-02	1.04E-01
Phylum	Proteobacteria	7.69E-05	4.89E-02	-1.23E-01	4.92E-03	5.70E-02	-1.34E-01
Order	Neisseriales	1.44E-04	1.83E-02	1.44E-02	6.53E-03	1.12E-01	1.02E-02
Family	Prevotellaceae	4.95E-04	2.38E-02	2.65E-02	7.87E-03	6.35E-02	2.43E-02
Family	Rhizobiaceae	4.29E-03	1.53E-02	-1.29E-02	5.57E-02	1.93E-02	-1.24E-02
Family	Veillonellaceae	8.96E-04	2.19E-02	1.77E-02	9.08E-03	5.97E-02	1.67E-02
Family	Neisseriaceae	1.44E-04	1.83E-02	1.44E-02	6.53E-03	1.12E-01	1.02E-02
Family	Carnobacteriaceae	4.29E-03	4.43E-02	1.11E-02	4.46E-02	8.11E-02	1.09E-02
Family	Fusobacteriaceae	4.95E-04	2.78E-02	9.46E-03	1.15E-02	8.45E-02	8.75E-03
Genus	Prevotella	4.95E-04	2.38E-02	2.65E-02	7.87E-03	6.35E-02	2.43E-02
Genus	Streptococcus	2.29E-03	4.44E-02	3.14E-02	1.02E-01	1.13E-01	3.04E-02
Genus	Fusobacterium	4.95E-04	2.78E-02	9.46E-03	1.15E-02	8.45E-02	8.75E-03
Otu	Corynebacterium kroppenstedtii	2.24E-06	1.06E-02	1.85E-02	2.26E-04	6.35E-02	1.78E-02
Otu	Corynebacterium sp.	6.62E-05	2.42E-02	1.29E-02	1.62E-03	1.18E-01	1.10E-02
Otu	Corynebacterium kroppenstedtii	1.06E-06	2.24E-02	2.22E-02	9.68E-04	6.35E-02	2.23E-02
Otu	Prevotella sp.	6.80E-03	2.03E-02	1.06E-02	7.17E-02	5.97E-02	9.20E-03

Table A.7: Significant taxa for PSOL-CTRL in the unmatched (U) but not significant in matched cohort (M). Wilx and Mas correspond to Wilcoxon ranked sum and MaAsLin respectively

Appendix B

Supplementary information for Chapter 6

This appendix contains supplemental data for the integrative network chapter.

Module	Description	ID	pvalue	p.adjust
black	extracellular matrix organization	GO:0030198	9.67E-18	1.42E-14
black	extracellular structure organization	GO:0043062	9.67E-18	1.42E-14
black	vasculature development	GO:0001944	3.64E-14	3.57E-11
black	cardiovascular system development	GO:0072358	4.88E-14	3.60E-11
black	blood vessel development	GO:0001568	1.86E-13	1.09E-10
blue	anterior/posterior pattern specification	GO:0009952	2.48E-07	9.22E-04
blue	regionalization	GO:0003002	5.22E-07	9.70E-04
blue	pattern specification process	GO:0007389	6.25E-06	7.75E-03
blue	cell part morphogenesis	GO:0032990	1.48E-05	1.38E-02
blue	cell projection morphogenesis	GO:0048858	2.07E-05	1.54E-02
brown	leukocyte activation	GO:0045321	1.97E-66	6.29E-63
brown	lymphocyte activation	GO:0046649	5.18E-65	8.27E-62
brown	leukocyte cell-cell adhesion	GO:0007159	8.61E-58	9.17E-55
brown	lymphocyte aggregation	GO:0071593	1.51E-56	1.21E-53
brown	T cell activation	GO:0042110	1.17E-55	6.21E-53
cyan	extracellular matrix organization	GO:0030198	2.40E-12	2.99E-09
cyan	extracellular structure organization	GO:0043062	2.40E-12	2.99E-09
cyan	blood vessel development	GO:0001568	3.62E-09	3.00E-06
cyan	angiogenesis	GO:0001525	5.05E-09	3.14E-06
cyan	vasculature development	GO:0001944	7.49E-09	3.65E-06
darkred	response to type I interferon	GO:0034340	2.87E-31	1.87E-28
darkred	innate immune response	GO:0045087	3.14E-31	1.87E-28
darkred	defense response to virus	GO:0051607	1.22E-30	4.83E-28

darkred	type I interferon signaling pathway	GO:0060337	1.28E-29	3.04E-27
darkred	cellular response to type I interferon	GO:0071357	1.28E-29	3.04E-27
green	ncRNA metabolic process	GO:0034660	5.57E-47	1.97E-43
green	ribonucleoprotein complex biogenesis	GO:0022613	2.69E-42	4.75E-39
green	ribosome biogenesis	GO:0042254	2.00E-41	2.36E-38
green	ncRNA processing	GO:0034470	4.33E-40	3.83E-37
green	RNA processing	GO:0006396	4.73E-34	3.35E-31
greenyellow	monocarboxylic acid metabolic process	GO:0032787	2.48E-19	4.81E-16
greenyellow	fatty acid metabolic process	GO:0006631	6.02E-18	5.84E-15
greenyellow	lipid biosynthetic process	GO:0008610	1.31E-14	8.47E-12
greenyellow	small molecule biosynthetic process	GO:0044283	5.67E-14	2.75E-11
greenyellow	acylglycerol metabolic process	GO:0006639	1.28E-12	4.97E-10
grey	detection of chemical stimulus involved in se	GO:0050911	2.71E-40	1.23E-36
grey	sensory perception of smell	GO:0007608	1.55E-37	3.49E-34
grey	detection of chemical stimulus involved in se	GO:0050907	1.11E-36	1.68E-33
grey	detection of stimulus involved in sensory per	GO:0050906	1.51E-31	1.71E-28
grey	sensory perception of chemical stimulus	GO:0007606	4.90E-31	4.43E-28
grey60	activation of immune response	GO:0002253	5.06E-15	9.95E-12
grey60	positive regulation of immune response	GO:0050778	2.33E-13	2.29E-10
grey60	innate immune response	GO:0045087	9.40E-12	6.16E-09
grey60	innate immune response-activating signal tran	GO:0002758	1.88E-10	8.49E-08
grey60	adaptive immune response	GO:0002250	2.92E-10	8.49E-08
lightcyan	nuclear-transcribed mRNA catabolic process, n	GO:0000184	2.01E-27	3.31E-24
lightcyan	SRP-dependent cotranslational protein targeti	GO:0006614	6.43E-25	5.31E-22
lightcyan	protein targeting to ER	GO:0045047	1.68E-24	8.83E-22
lightcyan	cotranslational protein targeting to membrane	GO:0006613	2.67E-24	8.83E-22
lightcyan	establishment of protein localization to endo	GO:0072599	2.67E-24	8.83E-22
lightgreen	molting cycle	GO:0042303	1.86E-13	9.20E-11
lightgreen	hair cycle	GO:0042633	1.86E-13	9.20E-11
lightgreen	epidermis development	GO:0008544	6.96E-09	2.29E-06
lightgreen	hair follicle development	GO:0001942	2.01E-05	3.10E-03
lightgreen	molting cycle process	GO:0022404	2.01E-05	3.10E-03
lightyellow	glycolipid metabolic process	GO:0006664	1.39E-03	4.46E-01
lightyellow	liposaccharide metabolic process	GO:1903509	1.55E-03	4.46E-01
lightyellow	membrane lipid metabolic process	GO:0006643	1.59E-03	4.46E-01
lightyellow	sphingolipid catabolic process	GO:0030149	2.40E-03	5.05E-01
lightyellow	membrane lipid catabolic process	GO:0046466	3.49E-03	5.87E-01
magenta	synapse assembly	GO:0007416	2.57E-06	6.73E-03
magenta	regulation of synapse assembly	GO:0051963	1.85E-05	1.62E-02
magenta	inorganic anion transmembrane transport	GO:0098661	2.15E-05	1.62E-02
magenta	inorganic ion transmembrane transport	GO:0098660	2.48E-05	1.62E-02

magenta	monovalent inorganic cation transport	GO:0015672	3.53E-05	1.75E-02
midnightblue	keratinization	GO:0031424	2.62E-08	5.55E-05
midnightblue	skin development	GO:0043588	7.59E-08	8.03E-05
midnightblue	keratinocyte differentiation	GO:0030216	4.46E-07	3.14E-04
midnightblue	regulation of water loss via skin	GO:0033561	5.22E-06	2.76E-03
midnightblue	epidermis development	GO:0008544	9.45E-06	3.41E-03
pink	anatomical structure regression	GO:0060033	1.12E-04	2.89E-01
pink	regulation of synaptic plasticity	GO:0048167	8.82E-04	5.01E-01
pink	apoptotic process involved in morphogenesis	GO:0060561	1.09E-03	5.01E-01
pink	chemical synaptic transmission	GO:0007268	1.39E-03	5.01E-01
pink	anterograde trans-synaptic signaling	GO:0098916	1.39E-03	5.01E-01
purple	inflammatory response	GO:0006954	5.58E-09	7.83E-06
purple	response to biotic stimulus	GO:0009607	9.34E-09	7.83E-06
purple	response to external biotic stimulus	GO:0043207	9.74E-09	7.83E-06
purple	response to other organism	GO:0051707	9.74E-09	7.83E-06
purple	innate immune response	GO:0045087	7.07E-08	4.55E-05
red	mitotic cell cycle	GO:0000278	3.56E-95	1.06E-91
red	mitotic cell cycle process	GO:1903047	7.25E-95	1.08E-91
red	chromosome organization	GO:0051276	3.69E-93	3.65E-90
red	nuclear division	GO:0000280	3.72E-73	2.76E-70
red	organelle fission	GO:0048285	5.21E-70	3.09E-67
royalblue	triglyceride metabolic process	GO:0006641	1.73E-06	7.88E-04
royalblue	gluconeogenesis	GO:0006094	2.08E-06	7.88E-04
royalblue	hexose biosynthetic process	GO:0019319	2.34E-06	7.88E-04
royalblue	acylglycerol metabolic process	GO:0006639	2.90E-06	7.88E-04
royalblue	neutral lipid metabolic process	GO:0006638	3.15E-06	7.88E-04
salmon	protein methylation	GO:0006479	3.42E-04	2.53E-01
salmon	protein alkylation	GO:0008213	3.42E-04	2.53E-01
salmon	macromolecule methylation	GO:0043414	1.07E-03	4.00E-01
salmon	histone H3-K9 methylation	GO:0051567	1.08E-03	4.00E-01
salmon	histone lysine methylation	GO:0034968	2.41E-03	5.84E-01
tan	keratinization	GO:0031424	7.36E-12	9.34E-09
tan	keratinocyte differentiation	GO:0030216	1.03E-11	9.34E-09
tan	epidermal cell differentiation	GO:0009913	1.43E-11	9.34E-09
tan	skin development	GO:0043588	1.77E-11	9.34E-09
tan	peptide cross-linking	GO:0018149	4.88E-11	2.05E-08
turquoise	RNA splicing	GO:0008380	2.36E-27	7.05E-24
turquoise	RNA processing	GO:0006396	9.51E-27	1.42E-23
turquoise	mRNA processing	GO:0006397	2.18E-25	2.17E-22
turquoise	mRNA metabolic process	GO:0016071	1.50E-21	1.12E-18
turquoise	RNA splicing, via transesterification reactio	GO:0000375	3.50E-21	1.50E-18

yellow	calcium-dependent cell-cell adhesion via plas	GO:0016339	3.50E-09	1.18E-05
yellow	homophilic cell adhesion via plasma membrane	GO:0007156	2.32E-07	3.94E-04
yellow	extracellular matrix organization	GO:0030198	1.47E-06	1.25E-03
yellow	extracellular structure organization	GO:0043062	1.47E-06	1.25E-03
yellow	glycosaminoglycan catabolic process	GO:0006027	4.43E-06	2.84E-03

Table B.1: Top 5 GO BP terms for each ADL module

Module	Description	ID	pvalue	p.adjust
black	extracellular matrix organization	GO:0030198	4.26E-19	7.39E-16
black	extracellular structure organization	GO:0043062	4.26E-19	7.39E-16
black	vasculature development	GO:0001944	1.58E-14	1.83E-11
black	cardiovascular system development	GO:0072358	2.32E-14	2.02E-11
black	blood vessel development	GO:0001568	3.47E-14	2.41E-11
brown	immune effector process	GO:0002252	1.87E-68	6.83E-65
brown	positive regulation of immune response	GO:0050778	4.99E-68	9.12E-65
brown	leukocyte activation	GO:0045321	1.08E-65	1.31E-62
brown	innate immune response	GO:0045087	1.63E-65	1.49E-62
brown	leukocyte cell-cell adhesion	GO:0007159	1.59E-63	1.16E-60
cyan	actin filament bundle assembly	GO:0051017	1.03E-06	1.19E-03
cyan	actin filament bundle organization	GO:0061572	1.31E-06	1.19E-03
cyan	reactive oxygen species metabolic process	GO:0072593	1.60E-06	1.19E-03
cyan	cell-substrate adhesion	GO:0031589	2.09E-06	1.19E-03
cyan	extracellular matrix organization	GO:0030198	4.58E-06	1.75E-03
darkgreen	muscle contraction	GO:0006936	5.20E-21	7.53E-18
darkgreen	muscle system process	GO:0003012	1.76E-20	1.28E-17
darkgreen	muscle cell differentiation	GO:0042692	7.53E-12	3.64E-09
darkgreen	muscle cell development	GO:0055001	1.53E-11	5.56E-09
darkgreen	smooth muscle contraction	GO:0006939	3.26E-11	9.43E-09
greenyellow	monocarboxylic acid metabolic process	GO:0032787	5.76E-37	2.13E-33
greenyellow	small molecule biosynthetic process	GO:0044283	4.56E-26	8.41E-23
greenyellow	organic acid catabolic process	GO:0016054	1.03E-25	1.26E-22
greenyellow	small molecule catabolic process	GO:0044282	1.05E-24	9.69E-22
greenyellow	fatty acid metabolic process	GO:0006631	1.79E-24	1.32E-21
grey	detection of chemical stimulus involved in se	GO:0050911	2.07E-31	9.39E-28
grey	sensory perception of smell	GO:0007608	1.86E-29	4.22E-26
grey	detection of chemical stimulus involved in se	GO:0050907	1.96E-26	2.96E-23
grey	sensory perception of chemical stimulus	GO:0007606	4.24E-24	4.81E-21
grey	detection of chemical stimulus	GO:0009593	4.07E-21	3.69E-18
lightgreen	molting cycle	GO:0042303	4.75E-18	2.92E-15
lightgreen	hair cycle	GO:0042633	4.75E-18	2.92E-15

lightgreen	epidermis development	GO:0008544	1.16E-12	4.73E-10
lightgreen	hair follicle development	GO:0001942	1.03E-09	2.10E-07
lightgreen	molting cycle process	GO:0022404	1.03E-09	2.10E-07
magenta	tube formation	GO:0035148	5.26E-08	1.20E-04
magenta	stem cell differentiation	GO:0048863	7.15E-08	1.20E-04
magenta	epithelial tube formation	GO:0072175	2.44E-07	2.14E-04
magenta	tube development	GO:0035295	2.56E-07	2.14E-04
magenta	morphogenesis of embryonic epithelium	GO:0016331	8.64E-07	5.79E-04
red	mitotic cell cycle process	GO:1903047	2.12E-89	5.49E-86
red	mitotic cell cycle	GO:0000278	4.64E-87	6.01E-84
red	chromosome organization	GO:0051276	5.79E-76	5.00E-73
red	nuclear division	GO:0000280	1.05E-66	6.77E-64
red	organelle fission	GO:0048285	2.14E-65	1.11E-62
royalblue	response to peptide hormone	GO:0043434	8.38E-06	1.97E-02
royalblue	response to peptide	GO:1901652	2.03E-05	2.38E-02
royalblue	regulation of cellular carbohydrate metabolic	GO:0010675	3.22E-05	2.51E-02
royalblue	lipid catabolic process	GO:0016042	4.49E-05	2.63E-02
royalblue	negative regulation of protein phosphorylatio	GO:0001933	6.96E-05	3.08E-02
tan	skin development	GO:0043588	6.81E-20	1.74E-16
tan	keratinocyte differentiation	GO:0030216	1.93E-19	2.47E-16
tan	keratinization	GO:0031424	1.17E-17	9.99E-15
tan	epidermal cell differentiation	GO:0009913	8.11E-17	5.18E-14
tan	epidermis development	GO:0008544	2.15E-16	1.10E-13
turquoise	RNA processing	GO:0006396	1.89E-35	7.56E-32
turquoise	mRNA metabolic process	GO:0016071	2.13E-29	4.26E-26
turquoise	RNA splicing	GO:0008380	6.31E-22	8.42E-19
turquoise	mRNA processing	GO:0006397	8.13E-20	8.13E-17
turquoise	RNA splicing, via transesterification reactio	GO:0000375	6.37E-16	3.64E-13
yellow	extracellular matrix organization	GO:0030198	1.58E-12	1.70E-09
yellow	extracellular structure organization	GO:0043062	1.58E-12	1.70E-09
yellow	collagen fibril organization	GO:0030199	1.31E-11	9.42E-09
yellow	multicellular organismal macromolecule metabo	GO:0044259	2.32E-11	1.25E-08
yellow	multicellular organism metabolic process	GO:0044236	1.08E-10	4.65E-08

Table B.2: Top 5 GO BP terms for each ADNL module

Module	Description	ID	pvalue	p.adjust
black	vasculature development	GO:0001944	2.02E-18	4.84E-15
black	cardiovascular system development	GO:0072358	2.91E-18	4.84E-15
black	blood vessel development	GO:0001568	9.65E-18	1.07E-14
black	blood vessel morphogenesis	GO:0048514	7.38E-15	6.13E-12

black	angiogenesis	GO:0001525	4.79E-12	3.18E-09
darkgreen	muscle contraction	GO:0006936	9.31E-21	1.03E-17
darkgreen	muscle system process	GO:0003012	3.30E-19	1.83E-16
darkgreen	smooth muscle contraction	GO:0006939	2.53E-11	9.37E-09
darkgreen	myofibril assembly	GO:0030239	6.72E-10	1.86E-07
darkgreen	striated muscle cell development	GO:0055002	1.26E-09	2.80E-07
greenyellow	monocarboxylic acid metabolic process	GO:0032787	4.06E-30	1.57E-26
greenyellow	small molecule biosynthetic process	GO:0044283	1.66E-23	3.20E-20
greenyellow	organic acid catabolic process	GO:0016054	2.54E-20	2.51E-17
greenyellow	fatty acid metabolic process	GO:0006631	2.60E-20	2.51E-17
greenyellow	cofactor metabolic process	GO:0051186	5.47E-19	4.23E-16
grey	detection of chemical stimulus involved in se	GO:0050911	2.07E-28	9.41E-25
grey	sensory perception of smell	GO:0007608	3.69E-26	8.37E-23
grey	detection of chemical stimulus involved in se	GO:0050907	3.61E-25	5.46E-22
grey	sensory perception of chemical stimulus	GO:0007606	2.37E-23	2.69E-20
grey	detection of chemical stimulus	GO:0009593	8.68E-19	7.89E-16
lightcyan	cilium assembly	GO:0042384	5.80E-07	7.94E-04
lightcyan	cilium organization	GO:0044782	1.06E-06	7.94E-04
lightcyan	cilium morphogenesis	GO:0060271	1.42E-06	7.94E-04
lightcyan	organelle assembly	GO:0070925	5.61E-06	2.36E-03
lightcyan	cellular component assembly involved in morph	GO:0010927	1.90E-05	6.40E-03
lightgreen	molting cycle	GO:0042303	4.59E-13	4.13E-10
lightgreen	hair cycle	GO:0042633	4.59E-13	4.13E-10
lightgreen	epidermis development	GO:0008544	1.84E-08	1.11E-05
lightgreen	hair follicle morphogenesis	GO:0031069	9.13E-06	4.11E-03
lightgreen	epidermis morphogenesis	GO:0048730	1.51E-05	5.45E-03
magenta	dopaminergic neuron differentiation	GO:0071542	2.09E-07	7.39E-04
magenta	anion transmembrane transport	GO:0098656	7.77E-07	9.94E-04
magenta	sodium ion transport	GO:0006814	8.43E-07	9.94E-04
magenta	monovalent inorganic cation homeostasis	GO:0055067	3.18E-06	2.82E-03
magenta	sodium ion transmembrane transport	GO:0035725	8.97E-06	6.35E-03
salmon	skin development	GO:0043588	3.40E-20	1.13E-16
salmon	epidermis development	GO:0008544	7.58E-20	1.26E-16
salmon	keratinocyte differentiation	GO:0030216	1.21E-18	1.34E-15
salmon	epidermal cell differentiation	GO:0009913	2.81E-16	2.34E-13
salmon	keratinization	GO:0031424	5.78E-13	3.86E-10
turquoise	RNA processing	GO:0006396	7.39E-35	2.71E-31
turquoise	mRNA metabolic process	GO:0016071	2.28E-27	4.18E-24
turquoise	RNA splicing	GO:0008380	4.00E-23	4.89E-20
turquoise	mRNA processing	GO:0006397	1.14E-19	1.05E-16
turquoise	RNA splicing, via transesterification reactio	GO:0000375	4.35E-15	2.28E-12

yellow	extracellular matrix organization	GO:0030198	4.13E-24	7.95E-21
yellow	extracellular structure organization	GO:0043062	4.13E-24	7.95E-21
yellow	regulation of cell motility	GO:2000145	4.32E-12	4.71E-09
yellow	regulation of cell migration	GO:0030334	4.89E-12	4.71E-09
yellow	collagen fibril organization	GO:0030199	7.91E-12	5.23E-09

Table B.3: Top 5 GO BP terms for each CAD module

Module	Description	ID	pvalue	p.adjust
black	extracellular matrix organization	GO:0030198	9.16E-33	1.38E-29
black	extracellular structure organization	GO:0043062	9.16E-33	1.38E-29
black	vasculature development	GO:0001944	2.70E-19	2.72E-16
black	cardiovascular system development	GO:0072358	3.85E-19	2.91E-16
black	blood vessel development	GO:0001568	1.67E-18	1.01E-15
blue	skin development	GO:0043588	8.92E-08	3.17E-04
blue	epidermis development	GO:0008544	1.03E-06	1.82E-03
blue	inositol lipid-mediated signaling	GO:0048017	4.54E-05	5.38E-02
blue	phosphatidylinositol 3-kinase signaling	GO:0014065	8.28E-05	5.61E-02
blue	cellular lipid catabolic process	GO:0044242	8.42E-05	5.61E-02
brown	leukocyte activation	GO:0045321	7.46E-70	2.06E-66
brown	lymphocyte activation	GO:0046649	7.13E-67	9.86E-64
brown	leukocyte cell-cell adhesion	GO:0007159	1.01E-58	9.31E-56
brown	T cell activation	GO:0042110	8.31E-57	4.60E-54
brown	T cell aggregation	GO:0070489	8.31E-57	4.60E-54
darkgreen	muscle contraction	GO:0006936	1.46E-24	1.49E-21
darkgreen	muscle system process	GO:0003012	6.50E-23	3.33E-20
darkgreen	smooth muscle contraction	GO:0006939	3.21E-12	1.10E-09
darkgreen	actin filament-based process	GO:0030029	1.09E-10	2.78E-08
darkgreen	regulation of muscle contraction	GO:0006937	2.30E-10	4.71E-08
darkturquoise	B cell proliferation	GO:0042100	1.40E-04	6.87E-03
darkturquoise	humoral immune response	GO:0006959	3.86E-04	9.45E-03
darkturquoise	B cell activation	GO:0042113	7.52E-04	1.09E-02
darkturquoise	lymphocyte proliferation	GO:0046651	1.15E-03	1.09E-02
darkturquoise	mononuclear cell proliferation	GO:0032943	1.18E-03	1.09E-02
green	ncRNA metabolic process	GO:0034660	5.97E-23	2.51E-19
green	ribosome biogenesis	GO:0042254	1.34E-17	2.05E-14
green	ncRNA processing	GO:0034470	1.46E-17	2.05E-14
green	ribonucleoprotein complex biogenesis	GO:0022613	3.22E-16	3.38E-13
green	response to molecule of bacterial origin	GO:0002237	6.17E-15	5.19E-12
greenyellow	fatty acid metabolic process	GO:0006631	4.74E-17	7.55E-14
greenyellow	monocarboxylic acid metabolic process	GO:0032787	1.54E-16	1.23E-13

greenyellow	lipid biosynthetic process	GO:0008610	1.69E-12	8.97E-10
greenyellow	acylglycerol metabolic process	GO:0006639	4.20E-12	1.59E-09
greenyellow	neutral lipid metabolic process	GO:0006638	5.00E-12	1.59E-09
grey	sensory perception of smell	GO:0007608	2.23E-33	1.01E-29
grey	detection of chemical stimulus involved in se	GO:0050911	4.24E-32	9.60E-29
grey	detection of chemical stimulus involved in se	GO:0050907	3.04E-27	4.59E-24
grey	sensory perception	GO:0007600	4.79E-25	5.43E-22
grey	sensory perception of chemical stimulus	GO:0007606	6.91E-25	6.27E-22
lightgreen	molting cycle	GO:0042303	6.51E-16	1.28E-13
lightgreen	hair cycle	GO:0042633	6.51E-16	1.28E-13
lightgreen	epidermis development	GO:0008544	2.16E-09	2.85E-07
lightgreen	aging	GO:0007568	4.96E-07	4.90E-05
lightgreen	hair follicle development	GO:0001942	8.88E-06	4.75E-04
magenta	growth	GO:0040007	2.82E-06	7.04E-03
magenta	regulation of growth	GO:0040008	5.94E-06	7.04E-03
magenta	negative regulation of locomotion	GO:0040013	8.07E-06	7.04E-03
magenta	negative regulation of cellular component mov	GO:0051271	8.84E-06	7.04E-03
magenta	signal release	GO:0023061	1.28E-05	8.18E-03
purple	defense response to virus	GO:0051607	9.37E-43	1.93E-39
purple	response to type I interferon	GO:0034340	2.35E-41	2.42E-38
purple	type I interferon signaling pathway	GO:0060337	4.40E-40	2.26E-37
purple	cellular response to type I interferon	GO:0071357	4.40E-40	2.26E-37
purple	response to virus	GO:0009615	1.34E-38	5.50E-36
red	mitotic cell cycle process	GO:1903047	7.36E-115	1.56E-111
red	mitotic cell cycle	GO:0000278	1.00E-113	1.06E-110
red	chromosome organization	GO:0051276	1.17E-108	8.28E-106
red	nuclear division	GO:0000280	4.34E-93	2.29E-90
red	organelle fission	GO:0048285	6.40E-90	2.71E-87
salmon	cilium assembly	GO:0042384	1.40E-08	3.56E-05
salmon	cilium organization	GO:0044782	3.65E-08	3.85E-05
salmon	cilium morphogenesis	GO:0060271	5.73E-08	3.85E-05
salmon	cell projection assembly	GO:0030031	6.06E-08	3.85E-05
salmon	cellular component assembly involved in morph	GO:0010927	1.26E-07	6.38E-05
tan	peptide cross-linking	GO:0018149	5.94E-13	6.68E-10
tan	keratinization	GO:0031424	6.41E-12	3.60E-09
tan	keratinocyte differentiation	GO:0030216	2.26E-11	8.48E-09
tan	epidermal cell differentiation	GO:0009913	6.74E-10	1.90E-07
tan	skin development	GO:0043588	3.34E-08	7.52E-06
turquoise	RNA processing	GO:0006396	1.81E-36	5.62E-33
turquoise	mRNA metabolic process	GO:0016071	8.94E-31	1.38E-27
turquoise	RNA splicing	GO:0008380	4.41E-22	4.55E-19

turquoise	mRNA processing	GO:0006397	1.38E-18	1.07E-15
turquoise	RNA splicing, via transesterification reactio	GO:0000375	1.68E-15	7.45E-13
yellow	extracellular matrix organization	GO:0030198	8.18E-13	1.29E-09
yellow	extracellular structure organization	GO:0043062	8.18E-13	1.29E-09
yellow	regulation of endothelial cell migration	GO:0010594	1.28E-08	1.35E-05
yellow	vasculature development	GO:0001944	3.59E-08	2.00E-05
yellow	blood vessel development	GO:0001568	4.05E-08	2.00E-05

Table B.4: Top 5 GO BP terms for each PSOL module

Module	Description	ID	pvalue	p.adjust
black	lipid catabolic process	GO:0016042	2.32E-06	4.10E-03
black	monocarboxylic acid metabolic process	GO:0032787	4.64E-06	4.10E-03
black	response to peptide hormone	GO:0043434	8.78E-06	4.23E-03
black	fatty acid metabolic process	GO:0006631	9.93E-06	4.23E-03
black	lipid localization	GO:0010876	1.20E-05	4.23E-03
brown	leukocyte activation	GO:0045321	5.47E-60	1.20E-56
brown	lymphocyte activation	GO:0046649	3.65E-58	4.00E-55
brown	leukocyte cell-cell adhesion	GO:0007159	1.19E-50	8.70E-48
brown	T cell activation	GO:0042110	2.70E-49	1.18E-46
brown	T cell aggregation	GO:0070489	2.70E-49	1.18E-46
cyan	muscle contraction	GO:0006936	2.14E-30	3.75E-27
cyan	muscle system process	GO:0003012	4.77E-29	4.18E-26
cyan	regulation of muscle system process	GO:0090257	2.71E-16	1.58E-13
cyan	regulation of muscle contraction	GO:0006937	1.30E-15	5.70E-13
cyan	regulation of system process	GO:0044057	1.04E-13	3.65E-11
greenyellow	monocarboxylic acid metabolic process	GO:0032787	1.05E-35	3.95E-32
greenyellow	small molecule biosynthetic process	GO:0044283	5.02E-30	9.44E-27
greenyellow	organic acid catabolic process	GO:0016054	6.87E-27	8.62E-24
greenyellow	fatty acid metabolic process	GO:0006631	1.72E-25	1.62E-22
greenyellow	cofactor metabolic process	GO:0051186	1.54E-24	1.03E-21
grey	detection of chemical stimulus involved in se	GO:0050911	1.08E-22	4.90E-19
grey	sensory perception of smell	GO:0007608	4.44E-21	1.01E-17
grey	sensory perception of chemical stimulus	GO:0007606	2.28E-20	3.45E-17
grey	detection of chemical stimulus involved in se	GO:0050907	3.05E-20	3.46E-17
grey	detection of stimulus involved in sensory per	GO:0050906	2.01E-16	1.82E-13
lightgreen	molting cycle	GO:0042303	6.65E-14	3.47E-11
lightgreen	hair cycle	GO:0042633	6.65E-14	3.47E-11
lightgreen	epidermis development	GO:0008544	7.96E-10	2.77E-07
lightgreen	hair follicle development	GO:0001942	3.93E-06	6.54E-04
lightgreen	molting cycle process	GO:0022404	3.93E-06	6.54E-04

magenta	negative regulation of cell differentiation	GO:0045596	3.63E-08	1.31E-04
magenta	negative regulation of developmental process	GO:0051093	1.84E-07	3.33E-04
magenta	tube formation	GO:0035148	1.09E-06	1.32E-03
magenta	negative regulation of cell migration	GO:0030336	1.97E-06	1.78E-03
magenta	negative regulation of cell motility	GO:2000146	3.31E-06	2.40E-03
red	mitotic cell cycle process	GO:1903047	4.02E-85	5.09E-82
red	mitotic cell cycle	GO:0000278	6.30E-83	4.00E-80
red	chromosome organization	GO:0051276	5.60E-71	2.37E-68
red	chromosome segregation	GO:0007059	1.96E-66	6.23E-64
red	nuclear division	GO:0000280	6.97E-66	1.77E-63
salmon	organelle assembly	GO:0070925	1.70E-09	3.75E-06
salmon	mitotic cell cycle	GO:0000278	6.77E-09	5.37E-06
salmon	sister chromatid cohesion	GO:0007062	7.31E-09	5.37E-06
salmon	nuclear chromosome segregation	GO:0098813	1.46E-08	5.53E-06
salmon	chromosome organization	GO:0051276	1.58E-08	5.53E-06
turquoise	RNA processing	GO:0006396	2.15E-38	6.87E-35
turquoise	mRNA metabolic process	GO:0016071	1.24E-30	1.98E-27
turquoise	RNA splicing	GO:0008380	2.34E-25	2.49E-22
turquoise	mRNA processing	GO:0006397	2.63E-23	2.11E-20
turquoise	RNA splicing, via transesterification reactio	GO:0000375	4.98E-20	2.28E-17
yellow	extracellular matrix organization	GO:0030198	8.05E-34	1.58E-30
yellow	extracellular structure organization	GO:0043062	8.05E-34	1.58E-30
yellow	vasculature development	GO:0001944	9.05E-23	1.19E-19
yellow	cardiovascular system development	GO:0072358	1.75E-22	1.72E-19
yellow	blood vessel development	GO:0001568	2.49E-22	1.96E-19

Table B.5: Top 5 GO BP terms for each PSONL module

Module	Description	ID	pvalue	p.adjust
black	lipid localization	GO:0010876	1.88E-07	2.51E-04
black	monocarboxylic acid metabolic process	GO:0032787	2.94E-07	2.51E-04
black	glycerolipid metabolic process	GO:0046486	1.59E-06	8.31E-04
black	response to peptide	GO:1901652	2.37E-06	8.31E-04
black	lipid catabolic process	GO:0016042	2.43E-06	8.31E-04
darkgreen	muscle contraction	GO:0006936	6.60E-25	7.56E-22
darkgreen	muscle system process	GO:0003012	4.05E-23	2.32E-20
darkgreen	smooth muscle contraction	GO:0006939	2.99E-11	1.14E-08
darkgreen	actin filament-based process	GO:0030029	4.21E-11	1.21E-08
darkgreen	regulation of muscle system process	GO:0090257	6.06E-11	1.39E-08
greenyellow	monocarboxylic acid metabolic process	GO:0032787	2.60E-34	1.01E-30
greenyellow	small molecule biosynthetic process	GO:0044283	6.75E-30	1.31E-26

greenyellow	organic acid catabolic process	GO:0016054	6.19E-29	8.01E-26
greenyellow	small molecule catabolic process	GO:0044282	3.17E-26	3.08E-23
greenyellow	carboxylic acid catabolic process	GO:0046395	1.53E-25	1.19E-22
grey	detection of chemical stimulus involved in se	GO:0050911	1.45E-25	6.61E-22
grey	detection of chemical stimulus involved in se	GO:0050907	1.32E-23	2.99E-20
grey	sensory perception of smell	GO:0007608	2.39E-23	3.62E-20
grey	detection of stimulus involved in sensory per	GO:0050906	8.49E-19	9.65E-16
grey	sensory perception of chemical stimulus	GO:0007606	1.21E-18	1.10E-15
lightgreen	molting cycle	GO:0042303	1.30E-15	8.98E-13
lightgreen	hair cycle	GO:0042633	1.30E-15	8.98E-13
lightgreen	epidermis development	GO:0008544	3.66E-12	1.69E-09
lightgreen	skin development	GO:0043588	1.27E-07	3.71E-05
lightgreen	hair follicle development	GO:0001942	1.89E-07	3.71E-05
magenta	inorganic ion transmembrane transport	GO:0098660	1.95E-07	6.37E-04
magenta	anion transmembrane transport	GO:0098656	3.66E-07	6.37E-04
magenta	tube formation	GO:0035148	6.93E-07	8.05E-04
magenta	dopaminergic neuron differentiation	GO:0071542	2.16E-06	1.88E-03
magenta	inorganic anion transmembrane transport	GO:0098661	3.06E-06	2.13E-03
salmon	organelle assembly	GO:0070925	4.88E-07	5.92E-04
salmon	cilium assembly	GO:0042384	5.08E-07	5.92E-04
salmon	chromosome organization	GO:0051276	1.08E-06	6.65E-04
salmon	cilium organization	GO:0044782	1.14E-06	6.65E-04
salmon	cilium morphogenesis	GO:0060271	1.67E-06	7.78E-04
tan	skin development	GO:0043588	3.21E-21	1.17E-17
tan	keratinocyte differentiation	GO:0030216	4.92E-20	8.95E-17
tan	epidermis development	GO:0008544	7.89E-19	9.58E-16
tan	epidermal cell differentiation	GO:0009913	4.55E-18	4.14E-15
tan	keratinization	GO:0031424	1.30E-15	9.45E-13
turquoise	RNA processing	GO:0006396	1.84E-28	6.60E-25
turquoise	mRNA metabolic process	GO:0016071	2.60E-21	4.66E-18
turquoise	RNA splicing	GO:0008380	5.30E-21	6.33E-18
turquoise	mRNA processing	GO:0006397	4.95E-18	4.44E-15
turquoise	RNA splicing, via transesterification reactio	GO:0000375	1.23E-14	6.28E-12
yellow	extracellular matrix organization	GO:0030198	3.87E-32	7.91E-29
yellow	extracellular structure organization	GO:0043062	3.87E-32	7.91E-29
yellow	vasculature development	GO:0001944	4.04E-22	5.50E-19
yellow	blood vessel development	GO:0001568	6.78E-22	6.73E-19
yellow	cardiovascular system development	GO:0072358	8.24E-22	6.73E-19

Table B.6: Top 5 GO BP terms for each CPSO module

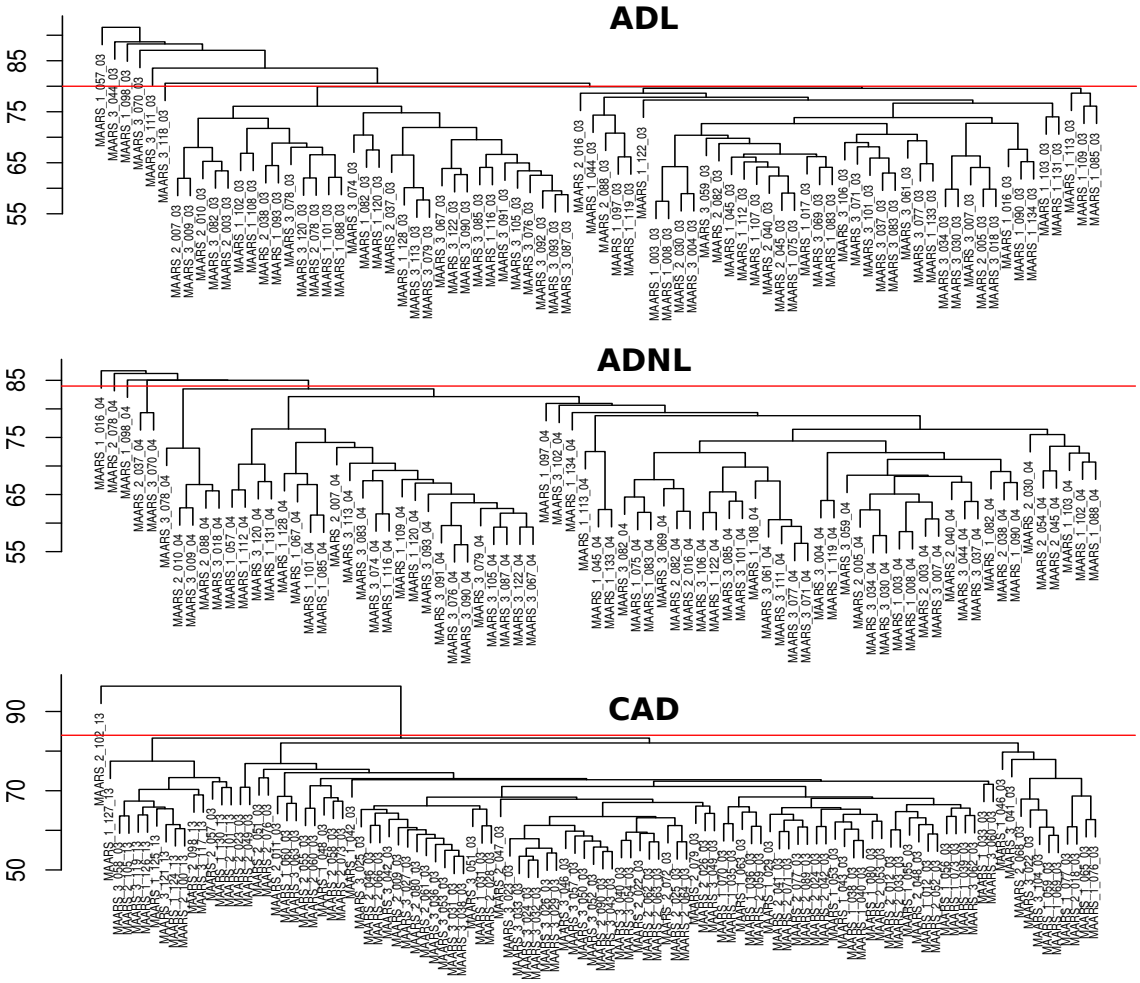

Figure B.1: Hierarchical clustering of ADL samples. Samples above the red line were excluded from network construction.

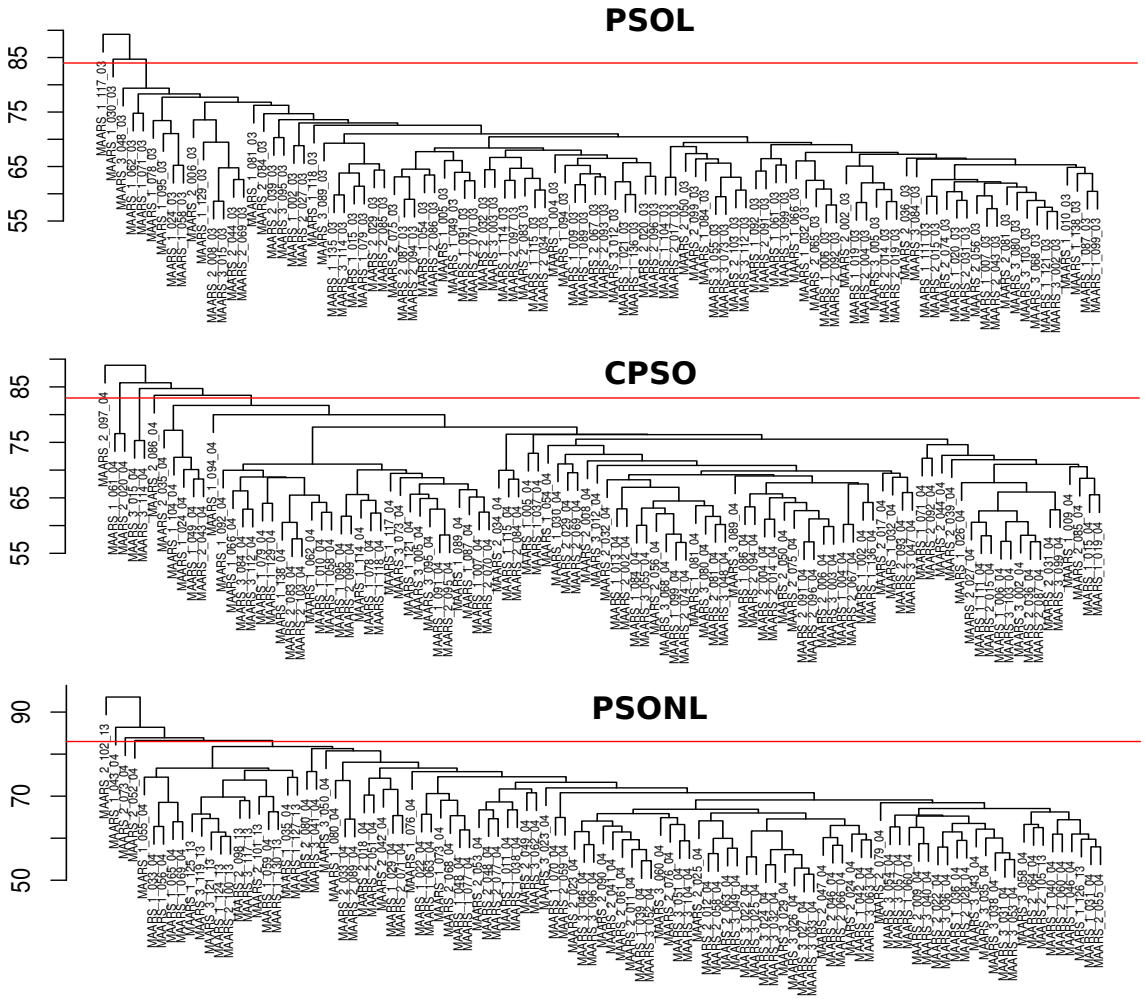

Figure B.2: Hierarchical clustering of PSOL samples. Samples above the red line were excluded from network construction.

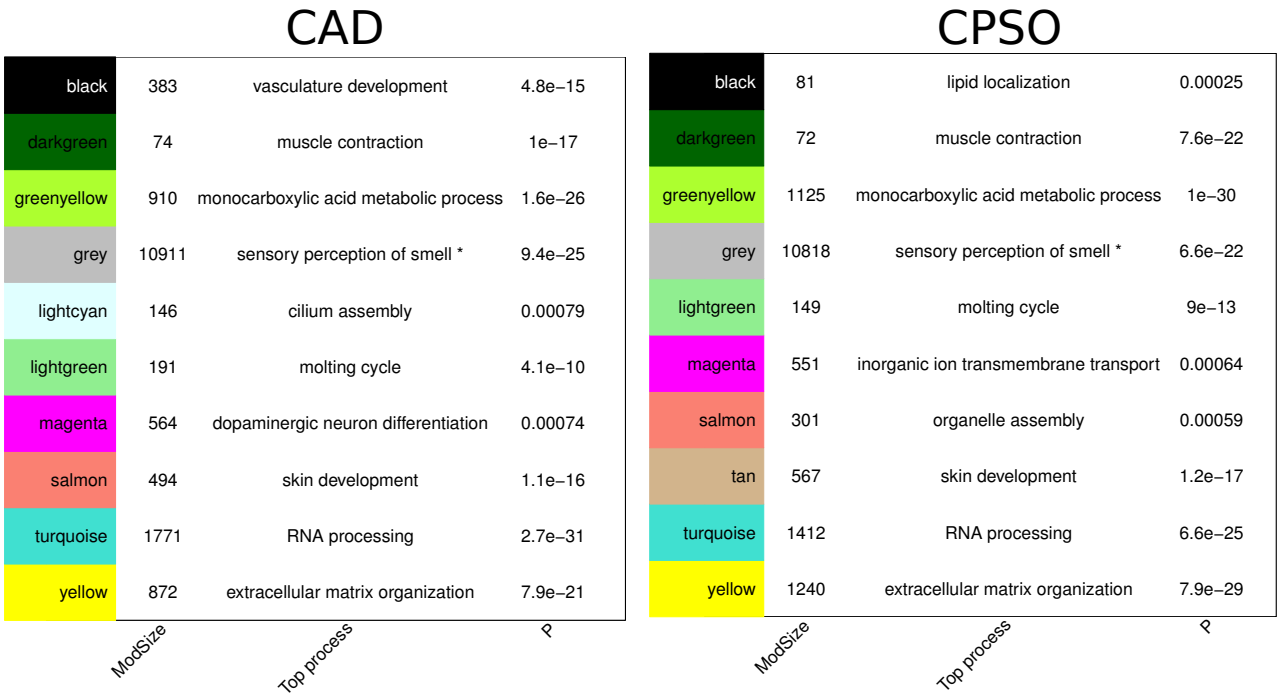

Figure B.3: Module definitions of Control co-expression networks. (Left) Module size and top GO biological process term for each module in the CAD network. (Right) Module size and GO terms for the CPSO network.

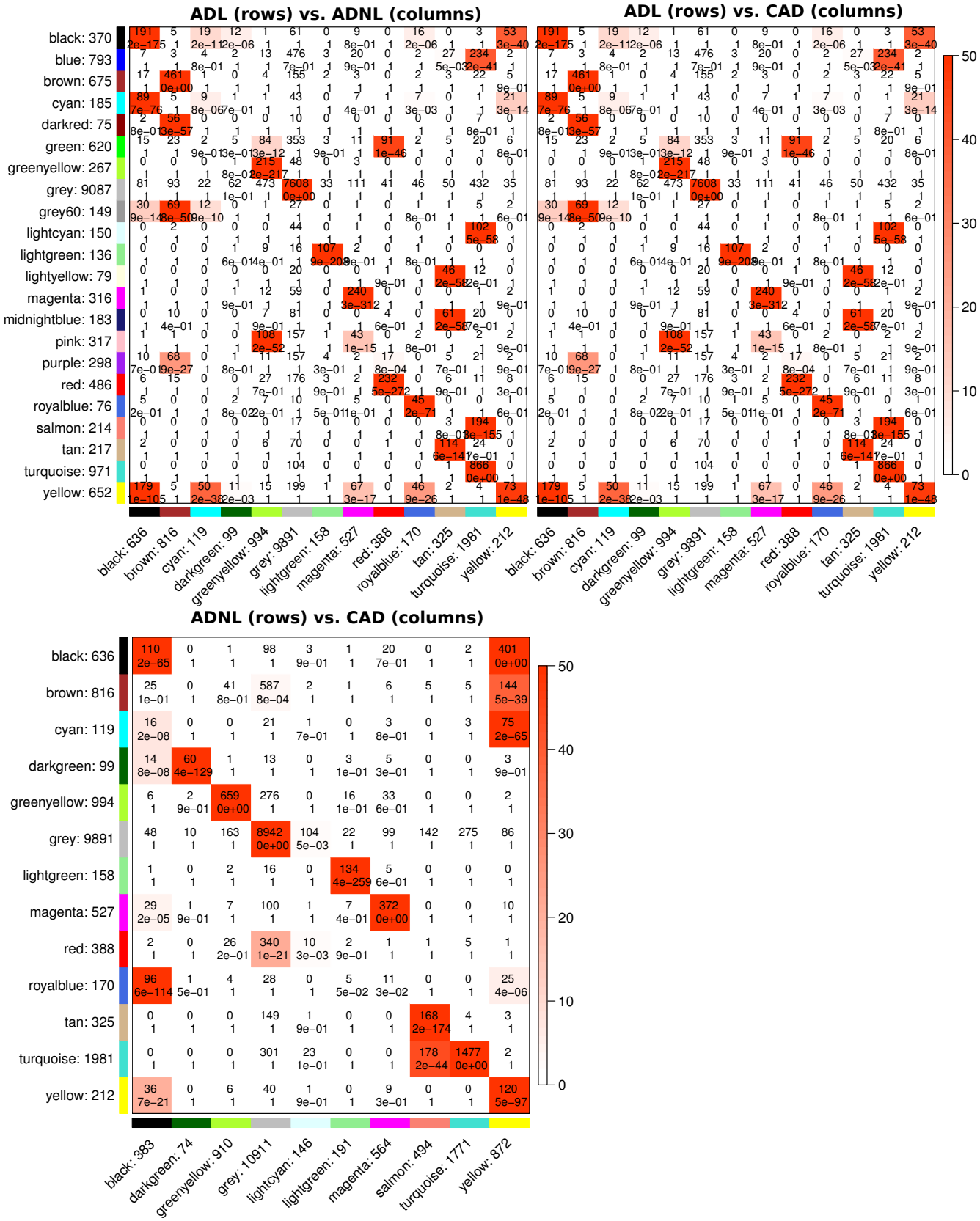

Figure B.4: Overlap of gene modules across AD networks. (Top left) - ADL modules vs ADNL modules. (Top-right) - ADL vs CAD. (Bottom-left) ADNL vs CAD. The first number in the heatmap tile is the number of genes which is present in both modules from both networks. The second number is the (uncorrected) Fisher's exact test p value. Heatmap tile color corresponds to -log(pvalue). P values were capped at -log(1e-50).

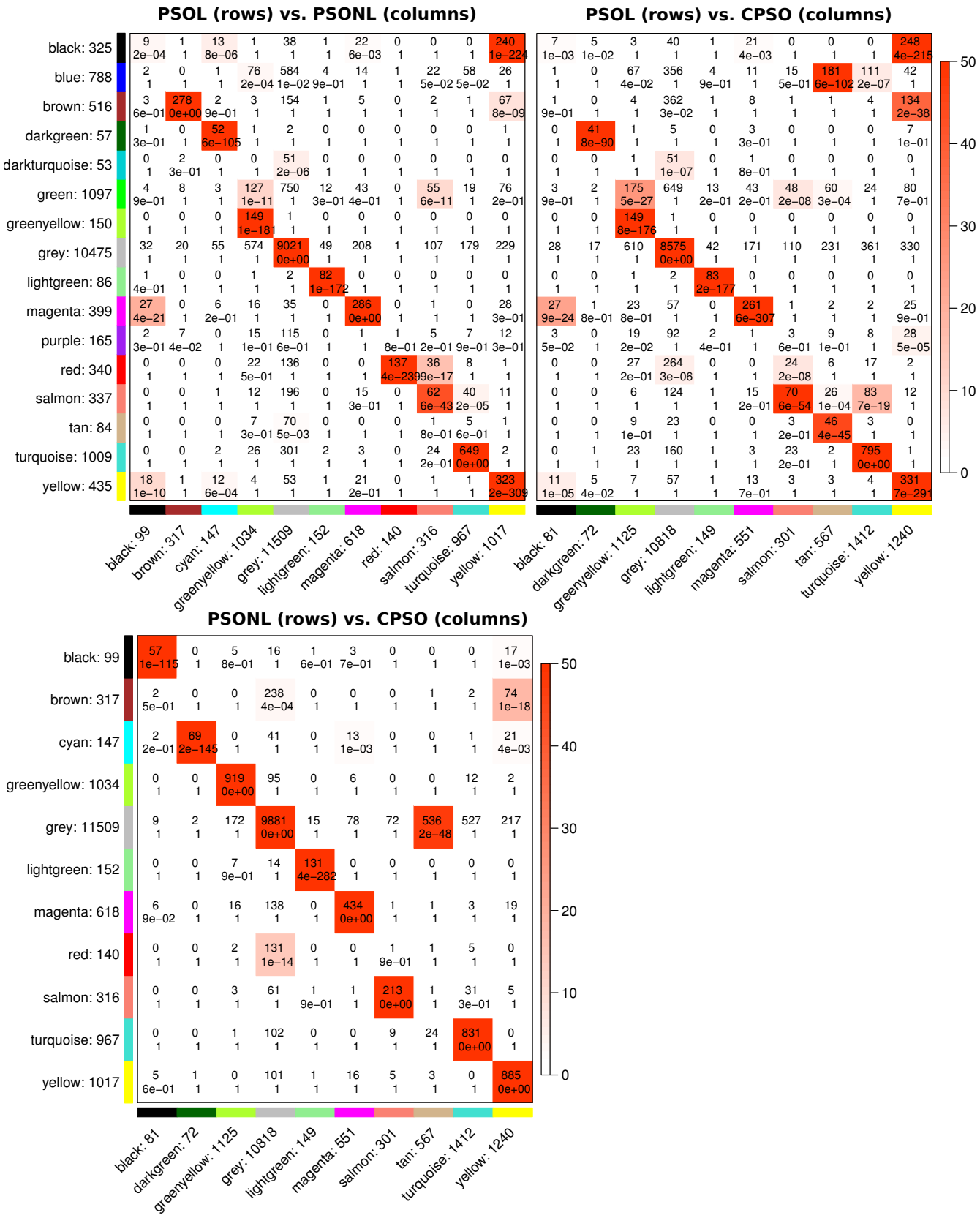

Figure B.5: Overlap of gene modules across PSO networks. (Top left) - PSOL modules vs PSONL modules. (Top-right) - PSOL vs CPSO. (Bottom-left) PSONL vs CPSO. The first number in the heatmap tile is the number of genes which is present in both modules from both networks. The second number is the (uncorrected) Fisher's exact test p value. Heatmap tile color corresponds to -log(pvalue). P values were capped at -log(1e-50).

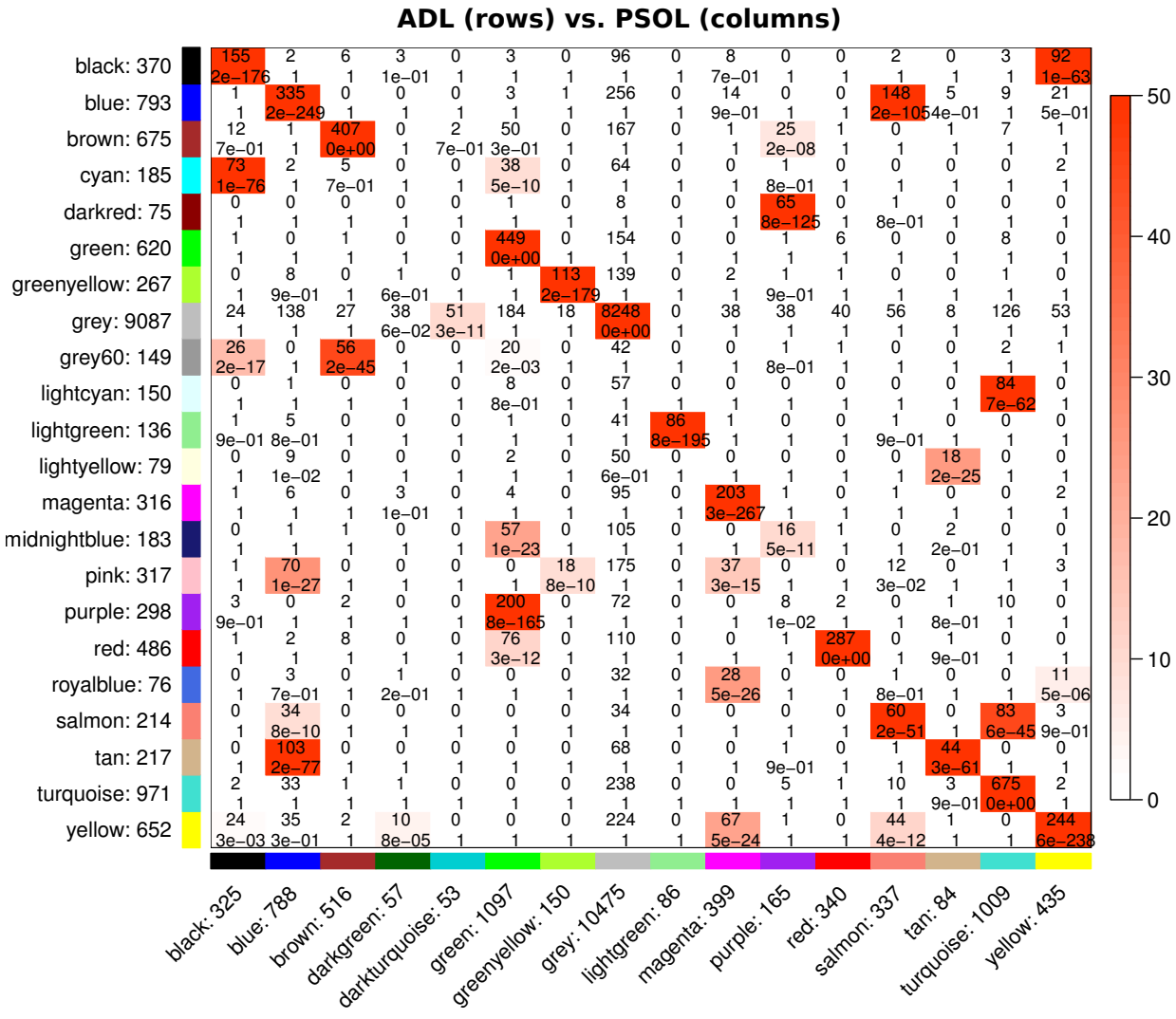

Figure B.6: Overlap of gene membership between ADL network modules and PSOL network modules. Rows correspond to AD modules, and columns correspond to PSOL modules. The first number in the heatmap tile is the number of genes which is present in both ADL and PSOL modules. The second number is the (uncorrected) Fisher's exact test p value. Heatmap tile color corresponds to -log(pvalue). P values were capped at -log(1e-50).

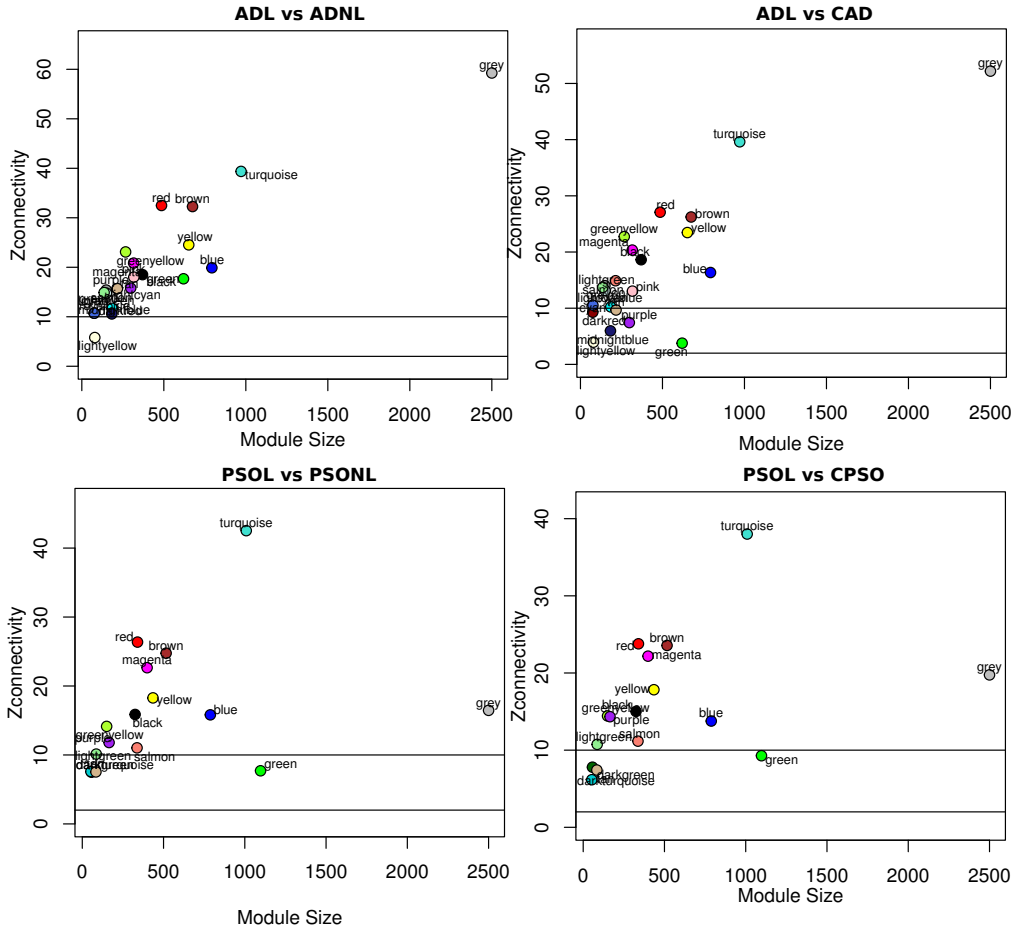

Figure B.7: Zconnectivity statistics for module preservation analysis. Zconnectivity is correlated with module size.

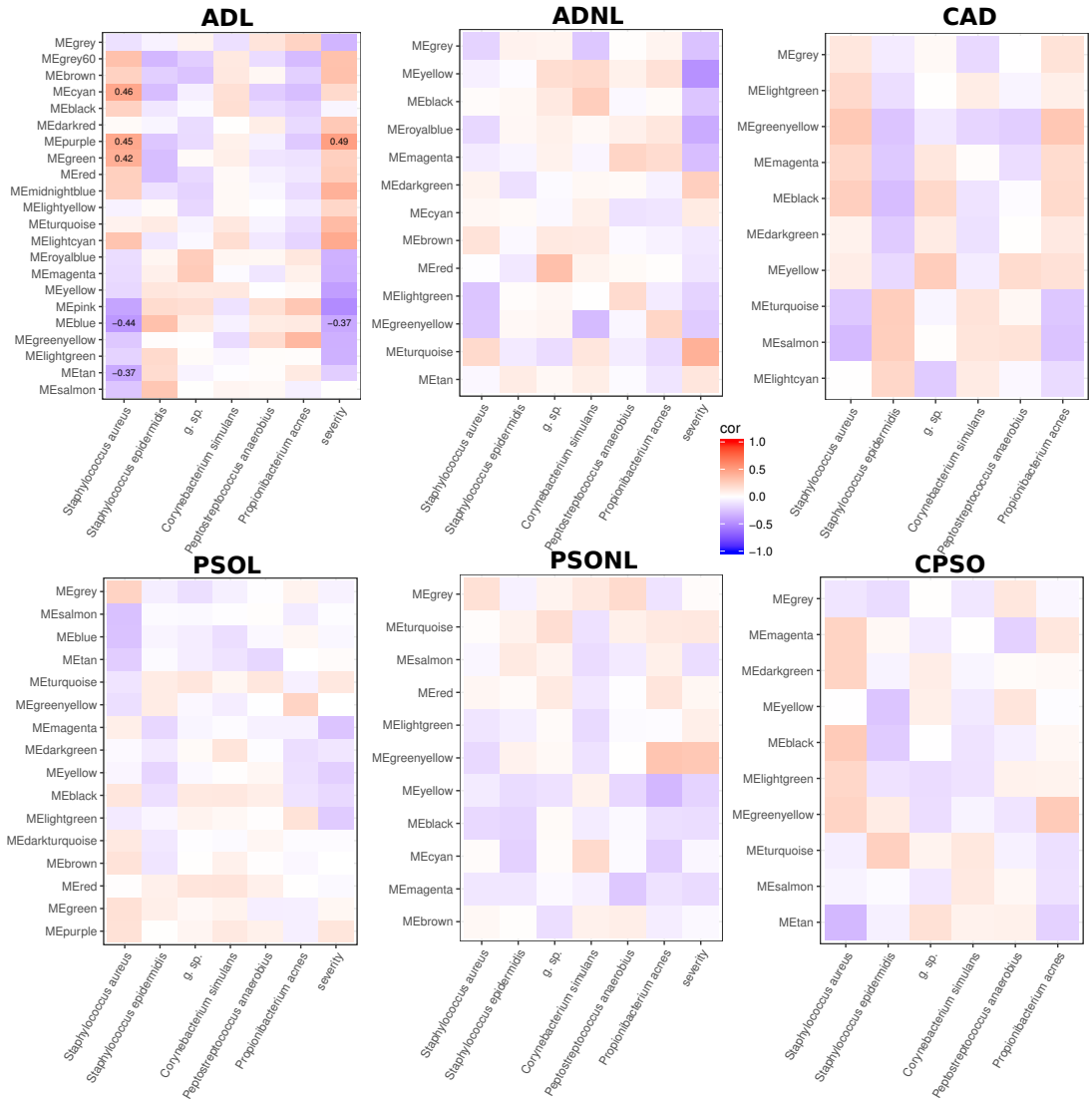

Figure B.8: Associations between microbes and module eigengenes across all networks. Heatmap tile corresponds to the correlation coefficient. Significant correlations ($p < 0.1$, $r > 0.30$) are marked.